UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Feature selection methods for classification with functional data

Máster Universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones (i^2-TIC) + Máster Universitario en Matemáticas y Aplicaciones (MMA)

Autor: RAMOS CARREÑO, Carlos
Tutor: BERRENDERO DÍAZ, José Ramón
SUÁREZ GONZALEZ, Alberto

FECHA: Septiembre, 2017

"I do not know what I may appear to the world, but to myself I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."

Isaac Newton

"Ce que nous connaissons est peu de chose, ce que nous ignorons est immense."

Pierre-Simon Laplace

Abstract

In this work, we present an extensive analysis of *Recursive Maxima Hunting* (RMH), which is a filter feature selection method for supervised classification problems with functional data.

In functional classification problems, the instances that are available for induction are characterized by a function of a continuous parameter rather than by a vector of attributes, as assumed by most standard machine learning methods. Functional data are intrinsically infinite dimensional and exhibit some structure associated to the assumed continuity and smoothness of the functions. Therefore, specially designed methods that employ the tools of Functional Data Analysis, the branch of statistics that deals with functional data, are needed.

The functional classification problem consists in discriminating trajectories that belong to two classes. In each of the classes, the trajectories are assumed to be realizations of a stochastic process with a different mean. Assuming homoscedasticity, both processes can be expressed as the sum of the corresponding mean and a common stochastic noise process $Z(t)$. In this context, a feature corresponds to some point in the trajectory. In general, the optimal classification rule depends on the whole trajectory. However, in many problems of interest, a finite, possibly small, subset of features can be sufficient to build accurate predictors. Feature selection consists in identifying such subsets. Given that functional data is infinite dimensional, this dimensionality reduction is important both for computational reasons and for the sake of interpretability. Moreover, in some cases, the optimal classification rule depends only in a finite number of features. In these cases, if the dimensionality reduction method preserves the features that appear in the optimal classification rule, the optimal classification could still be built after the reduction.

In Recursive Maxima Hunting (RMH), the feature selection method analyzed in this work, relevant features are identified in a iterative manner. Once a relevant feature has been identified, the trajectories are corrected by removing the information provided by the values of the trajectories at the corresponding location.

Initially, RMH selects the feature that maximizes the dependency of the value of the trajectory at that point and the class label. The information provided by this selected variable is then subtracted from all the sampled functions. This information is expressed as the conditional expectation of the noise process $Z(t)$ given the selected feature. The steps are then repeated until the dependencies between the class and each of the remaining unselected features are not significant. We show that the process of subtracting the information of the features selected reveals features that are not relevant by themselves, but are relevant in combination with the previously selected features.

From a complementary viewpoint, RMH can also be seen as providing an interpolation of the difference of the class means, based on the values of the trajectories at the selected points. The algorithm halts when the interpolation of the difference of means is sufficiently accurate. The form of this interpolation depends on the type of noise process $Z(t)$ assumed to compute the RMH corrections. For instance, if $Z(t)$ is a Brownian process, a linear interpolation between the values of the trajectory at the origin and at the selected points is made.

If the difference between the means is piecewise linear, and the noise assumed to compute the corrections is a Uniform-Brownian process, RMH selects the points

that appear in the optimal classification rule. The Uniform-Brownian process corresponds to the limit of an Ornstein-Uhlenbeck process whose lengthscale and variance tend to infinity, while the ratio of these two quantities remains constants.

Finally, we carry out an extensive empirical evaluation to compare several variants of RMH with other dimensionality reduction methods for functional classification. In the experiments synthetic datasets with different means and train sizes, and a Brownian noise process are used. The methods are tested using also real-world classification problems from different areas of application, which have previously been considered in the functional data literature. In both types of experiments, RMH achieves excellent overall results. Specifically, accurate classification is achieved in the problems analyzed using small numbers of selected features. We present also some examples that illustrate how RMH behaves when the noise process assumed in RMH does not coincide with the one used to generate the data, and when the class can be completely determined by the sampled functions.

Resumen

En este trabajo presentamos un amplio análisis de *Recursive Maxima Hunting* (RMH), un método de filtro para la selección de variables en problemas de clasificación supervisada con datos funcionales.

En los problemas de clasificación con datos funcionales, las muestras disponibles para la predicción están caracterizadas por una función de un parámetro continuo, en lugar de un vector de atributos, que es la suposición realizada en la mayoría de métodos estándar en aprendizaje automático. Los datos funcionales son intrínsecamente infinito-dimensionales, y exhiben una estructura asociada a la continuidad y suavidad que se asume a las funciones. Por tanto, se necesitan métodos especialmente diseñados que aprovechen las herramientas del Análisis de Datos Funcionales, la rama de la Estadística que trabaja con datos funcionales.

El problema de clasificación consiste en discriminar trayectorias que pertenecen a dos clases. En cada clase, supondremos que las trayectorias son realizaciones de un proceso estocástico con distinta media. Haciendo la hipótesis de homocedasticidad, ambos procesos pueden expresarse como la suma de la media correspondiente y un proceso de ruido común $Z(t)$. En este contexto, una variable corresponde a un punto en la trayectoria. En general, la regla de clasificación óptima depende de toda la trayectoria. No obstante, en muchos problemas de interés, un subconjunto de variables finito, posiblemente pequeño, puede ser suficiente para construir buenos predictores. La selección de variables consiste en identificar dichos subconjuntos. Dado que los datos funcionales tienen dimensión infinita, esta reducción de dimensionalidad es importante, tanto por razones de coste computacional como para facilitar la interpretación de los resultados. Además, en algunos casos, la regla de clasificación óptima depende solo en un número finito de variables. En esos casos, si mediante el método de reducción de dimensión se identifican las variables que aparecen en dicha regla óptima, sigue siendo posible conseguir la clasificación óptima después de la reducción de dimensionalidad.

En Recursive Maxima Hunting (RMH), el método de selección de variables analizado en este trabajo, las variables relevantes se identifican de forma iterativa. Una vez identificada una variable relevante, las trayectorias son corregidas de forma que se elimine la información proporcionada por los valores de las trayectorias para esa variable.

Inicialmente, RMH selecciona la variable que maximiza la dependencia entre el valor de la trayectoria en ese punto y la etiqueta de la clase. La información proporcionada por esta variable se sustrae para cada una de las trayectorias de la muestra. Esta información se expresa como la esperanza condicional del proceso de ruido $Z(t)$ dado el valor de la trayectoria en el punto seleccionado. Estos pasos se repiten hasta que las dependencias entre la clase y las variables sin seleccionar dejen de ser significativas. En este trabajo mostramos que el proceso de sustraer esta información de las variables seleccionadas permite identificar variables que no son relevantes por sí mismas, pero que sí lo son en combinación con las variables seleccionadas previamente.

Desde un punto de vista complementario, puede verse RMH como un método que proporciona una interpolación de la diferencia entre las medias de las dos clases, a partir de los valores de la trayectoria en los puntos seleccionados. El algoritmo finaliza cuando la interpolación de dicha diferencia es suficientemente precisa. La

forma de esta interpolación depende del tipo de proceso de ruido $Z(t)$ asumido para computar las correcciones. Por ejemplo, si $Z(t)$ es un proceso browniano, se realiza una interpolación lineal entre los puntos seleccionados.

Si la diferencia entre las medias es lineal a trozos, y el ruido asumido para calcular las correcciones viene dado por un proceso uniforme-browniano, RMH selecciona los puntos que aparecen en la regla de clasificación óptima. El proceso uniforme-browniano se define como el límite de un proceso Ornstein-Uhlenbeck cuya varianza y parámetro de escala tienden a infinito, mientras que su cociente se mantiene constante.

Finalmente, hemos realizado una amplia evaluación empírica, en la que se comparan variantes de RMH con otros métodos de reducción de dimensionalidad aplicables a clasificación funcional. En los experimentos se han usado conjuntos de datos sintéticos, con distintas medias y tamaños del conjunto de entrenamiento y un proceso de ruido browniano. También se han evaluado estos métodos usando problemas de clasificación con datos de problemas reales, que se presentan en distintas áreas de aplicación y que han sido considerados previamente en la literatura sobre datos funcionales. En ambos tipos de experimento, RMH obtiene buenos resultados en general. Concretamente, en los problemas analizados, se logra una clasificación precisa usando un número reducido de variables. También presentamos algunos ejemplos que ilustran el comportamiento de RMH cuando el proceso de ruido asumido por RMH no coincide con el usado para generar los datos, y cuando la etiqueta de clase se encuentra completamente determinada por las funciones de la muestra.

Acknowledgements

This Master thesis would not have been possible without the help and guidance of my advisors Dr. Alberto Suárez Gonzalez and Dr. José Ramón Berrendero Díaz. They provided me with a lot of useful suggestions to improve the quality of this thesis and my research work. I would like to give special thanks to Dr. José Luis Torrecilla Noguerales, who helped me to understand the RMH method and the previous work in the area, and also offered interesting suggestions for the thesis.

I would also like to extend my thanks to my colleagues at the Machine Learning Group of the Universidad Autónoma de Madrid, who also helped and supported me through the development of this work.

Finally, I wish to thank my family and friends, who always encouraged me to pursue my objectives. Without their support and understanding, finishing this work would have been much harder.

Contents

List of Figures

List of Tables

List of Abbreviations

dCor	**D**istance **Cor**relation
dCov	**D**istance **Cov**ariance
FDA	**F**unctional **D**ata **A**nalysis
GP	**G**aussian **P**rocess
k**-NN**	k **N**earest **N**eighbors
MH	**M**axima **H**unting
MI	**M**utual **I**nformation
mRMR	**M**inimum **R**edundancy **M**aximum **R**elevance
PCA	**P**rincipal **C**omponent **A**nalysis
PLS	**P**artial **L**east **S**quares
RKHS	**R**eproducing **K**ernel **H**ilbert **S**pace
RK-VS	**R**eproducing **K**ernel **V**ariable **S**election
RMH	**R**ecursive **M**axima **H**unting

List of Symbols

$\cdot \sim \cdot$	Equivalent measures
$\cdot \perp \cdot$	Mutually singular measures
$\mathrm{card}(\cdot)$	Cardinality of a set
$\mathrm{cdf}(\cdot)$	Cumulative distribution function of a standard normal random variable
$\mathcal{F}$	Feature space or σ-algebra, depending on the context
$g^*(\cdot)$	Optimal classifier (Bayes classifier)
$\mathcal{H}$	RKHS
$\mathcal{H}(K)$	RKHS associated with the kernel K
$I(\cdot, \cdot)$	Mutual information
$K(\cdot, \cdot)$	Covariance function (kernel) of a Gaussian Process
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
L	Classification error
L^*	Optimal classification error (Bayes error)
$\hat{L}_n$	Empirical risk
$\mathrm{pdf}(\cdot)$	Probability density function of a standard normal random variable
$\mathcal{R}(\cdot, \cdot)$	Distance correlation
$\mathcal{R}_n(\cdot, \cdot)$	Sample distance correlation
$\mathrm{Red}(\cdot)$	Redundancy
$\mathrm{Rel}(\cdot)$	Relevance
$\mathcal{V}(\cdot, \cdot)$	Distance covariance
$\mathcal{V}_n(\cdot, \cdot)$	Sample distance covariance
$X(\cdot)$	Stochastic process in the classification problem
Y	Dichotomic random variable representing class labels
$Z(\cdot)$	Noise stochastic process
$\eta(\cdot)$	Discriminant rule for the Bayes classifier
$\mu(\cdot)$	Mean function
$\sigma(\cdot)$	Standard deviation function
$\Omega_n(\cdot, \cdot)$	Unbiased estimator of the squared distance covariance

Chapter 1

Introduction

The objective of this thesis is to introduce and analyze *Recursive Maxima Hunting* (RMH), a new feature selection method in the context of Functional Data Analysis.

RMH is a filter method that can be applied to classification problems. It iteratively selects the non previously selected feature that maximizes the dependency with the class labels, using an appropriate dependency measure such as distance correlation. After each feature is selected, it removes from the data the information of the selected feature. That information is defined in this method as the conditional expectation of the noise stochastic process of the corresponding class, given that the value of the selected feature was observed.

Functional Data Analysis (FDA) is a branch of statistics in which the available data are functions of a continuous parameter, instead of real numbers or vectors. Although the first references of functional data date from the 1980s (Ramsay, 1982), the interest in the field began to increase in the 1990s, because the techniques to record and process functional data became more widely available. Since then, FDA has been gaining popularity, and now is a very active research field in statistics (Cuevas, 2014; Wang, Chiou, and Müller, 2016).

Functional data are intrinsically infinite dimensional and posses characteristics that set them apart from multivariate data. In particular, if the functions are continuous, as is often the case, there exists a strong correlation between the values of the function at nearby points. This is in contrast with classical multivariate statistics, where the data are finite dimensional and not necessarily correlated. A simple-minded approach to dealing with functional data is to discretize the observed function and model the resulting vector using the techniques of classical multivariate statistics. However, in this analysis, the information provided by the functional nature of the data is lost. It is preferable to develop specific statistical tools that take into account the functional structure of the data.

Nevertheless, some ideas in classical statistics have an equivalence when dealing with functional data. In classical statistics, the observed data are assumed to be a sample from a population that can be described by a probability distribution. This idea can be translated to functional data where the population can be described by a stochastic process. Gaussian and multivariate Gaussian distributions in classical statistics have their functional correlate in the family of Gaussian processes. These are stochastic processes whose marginals over an arbitrary selection of points are multivariate Gaussian distributions. Similarly to the multivariate Gaussian distribution, which is completely determined by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, a Gaussian process is completely determined by a mean function $\mu(t)$ and a covariance function, or kernel, $K(s,t)$.

In this thesis we will assume that the data are real-valued functions $f : [0,1] \to \mathbb{R}$. The continuous parameter on which the function depends will be referred to as t, s, r or u, as needed. This continuous parameter could be seen as a time, in which case the

corresponding function can be seen as a trajectory (i.e. a time series in continuous time).

In practice, these functions are discretized. They are represented by a vector of observations at a finite set of points $(f(t_0), f(t_1), \ldots, f(t_n))$. Nevertheless, the methods used to analyze these data will take advantage of the functional structure, assuming that the values of the function at nearby points are strongly correlated.

1.1 Classification problems with functional data

The goal of supervised classification is to predict the class label of an instance that is characterized by a set of features. To this end, we have at our disposal some labeled examples. These training data consist of a collection of pairs $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, in which $\mathbf{x}_n$ are the features that characterize the n-th example. The class of this example is $y_n \in \{C_1, \ldots, C_K\}$. When $K = 2$, the classification problem is *binary*. In most standard machine learning methods one assumes that the instances are described by a D dimensional input vector. Multivariate statistics techniques are then used to model the relationship between the input vector and the class label.

In the context of FDA, the instances are described by an input function instead of an input vector. One example of this type of problem is the Berkeley Growth Study (Ramsay, 2006; Mosler and Mozharovskyi, 2015). The data for these problem are shown in Figure 1.1. In the Berkeley Growth Study, the space of features is the functional space $\mathcal{F} = \{f : [1, 18] \to \mathbb{R}\}$, where f are functions that yield the heights of children measured at ages 1 to 18. The objective of this classification problem is to decide, given a curve, whether the growth profile belongs to a boy (class 0) or a girl (class 1). Our goal is to induce a hypothesis from the available data that can predict the gender of an individual on the basis of his or her growth profile. Mathematically, the hypothesis is a measurable function $g : \mathcal{F} \to \{0, 1\}$, which is called a *classifier* or *classification rule*.

In this work, we will restrict ourselves to binary classification problems, in which the functions to classify are sampled from two populations, P_0 and P_1. The instances are characterized by one-dimensional functions $X(t)$ where $t \in [0, 1]$. The class label Y is a dichotomic random variable takes the value 0 when X is sampled from P_0 and 1 when X is sampled from P_1.

To formulate the problem, we will express functions sampled from a particular population as the sum of a zero-mean stochastic noise process and a deterministic function, which corresponds to the mean,

$$X(t) = \begin{cases} \mu_0(t) + Z_0(t), & \text{if } Y = 0, \\ \mu_1(t) + Z_1(t), & \text{if } Y = 1. \end{cases}$$

In this expression $\mu_0(t) = \mathbb{E}[X(t)|Y = 0]$, $\mu_1(t) = \mathbb{E}[X(t)|Y = 1]$ denote the means (deterministic functions) and $Z_0(t), Z_1(t)$ the stochastic parts, for class 0 and class 1, respectively. The prior probability of the class being 1, $\mathbb{P}(Y = 1)$, is denoted as p. Thus, $\mathbb{P}(Y = 0) = 1 - p$.

Assuming that the mean of one of the processes is known, say $\mu_0(t)$, the problem is equivalent to discriminating between a zero mean stochastic process and and a general stochastic process

$$X(t) = \begin{cases} Z_0(t), & \text{if } Y = 0, \\ \mu(t) + Z_1(t), & \text{if } Y = 1, \end{cases}$$

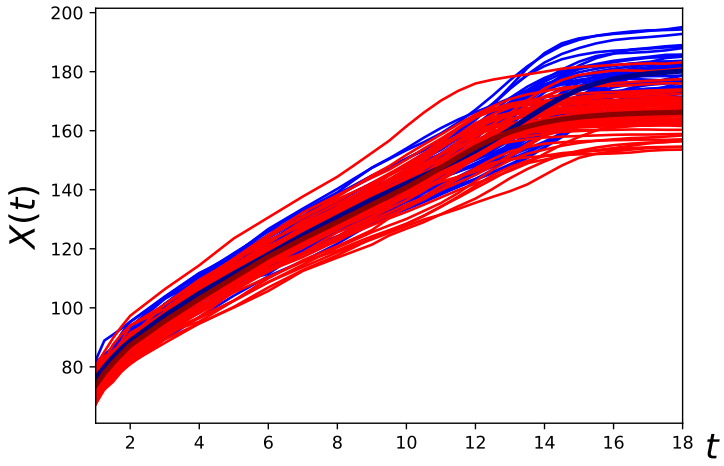

FIGURE 1.1: Dataset corresponding to the Berkeley Growth Study. In this study the heights of 39 boys (in blue) and 54 girls (in red) were monitored from age 1 to age 18. The binary classification problem consists in predicting the gender of an individual from her or his growth profile.

where $\mu(t) = \mu_1(t) - \mu_0(t)$. If $\mu_0(t)$ is not known, it can be estimated from the training data that belongs to class 0 and then subtracted from all trajectories. Therefore, without much loss of generality, we will focus on this second type of functional classification problem.

We will further restrict our attention to a homoscedastic setting, in which the stochastic noise process, $Z(t)$, is the same in both classes. Therefore, trajectories from these classes are distinguishable only by their means:

$$X(t) = \begin{cases} Z(t), & \text{if } Y = 0, \\ \mu(t) + Z(t), & \text{if } Y = 1. \end{cases} \tag{1.1}$$

Examples of these types of problems are given in Figure 1.2.

Because of the functional nature of the data, nearby locations in a given trajectory, which generally exhibit strong correlations, should convey similar information. These high levels of redundancy suggest that an the values of the trajectories at a set of appropriately selected locations can be sufficient to yield accurate class label predictions. In consequence, the goal of this work is to design a filter feature selection method to be used in conjunction with a standard classifier (e.g. nearest-neighbors) to reduce the dimensionality of the problem and improve the accuracy of the predictions. The process consists in selecting from the potentially infinite set of features available for discrimination (that is, the values of the trajectories at the different points), a finite subset that are jointly *relevant* to the classification task, and are minimally *redundant*.

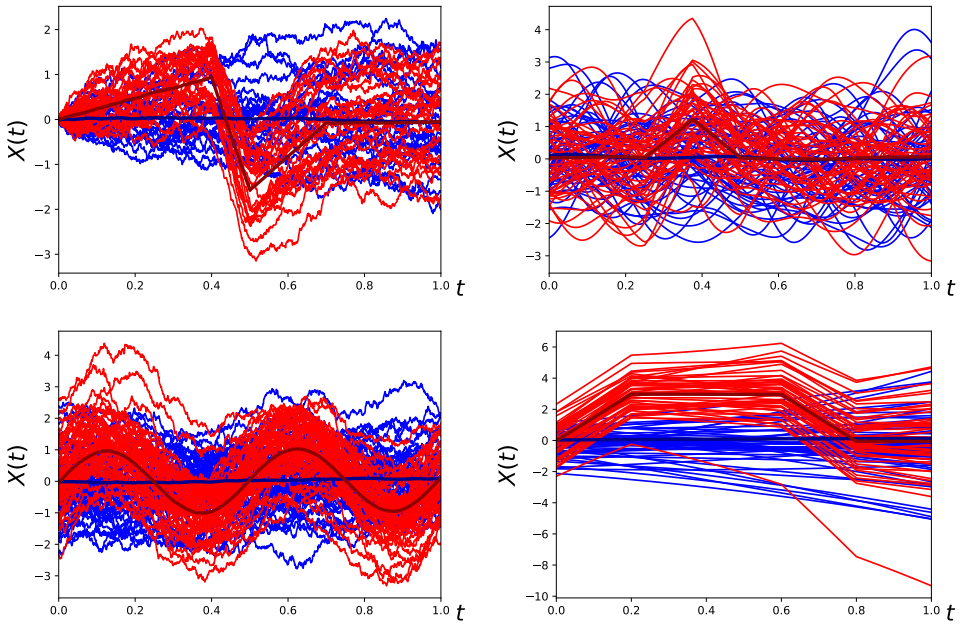

FIGURE 1.2: Trajectories for different functional classification problems that obey Equation (1.1). In the upper left plot, Z is a Brownian process and μ is a piecewise linear function. In the upper right plot, Z is a Radial Basis Function (RBF) Gaussian process and μ is a small peak. In the lower left plot, Z is an exponential Gaussian process and and μ is a sinusoidal function. In the lower right plot, Z is a polynomial Gaussian process and μ is another piecewise linear function. In all of these examples, the blue trajectories correspond to the case $Y = 0$ and the red ones are the trajectories where $Y = 1$. The widest lines are the respective means.

1.2 Dimensionality reduction

In classical multivariate statistics, it is common to have classification problems with a high number of features. If all these features were necessary, one would incur high storage costs. Furthermore, the process of training classifiers and their evaluation would be computationally costly. For example, a 64×64 small grayscale image that could be used in character recognition has 4096 features. For functional data, this problem can be more acute, because the number of features is infinite.

However, in some cases, it is possible to apply a *dimensionality reduction* method that identifies a smaller number of features, which could be a combination of the original ones, that are sufficient for the induction of accurate classifiers from the training data. Since these methods often discard redundant or irrelevant features, it is possible that classifiers trained on the reduced input space are more accurate that classifiers trained with all the original features.

1.2.1 Variable selection methods

Variable selection, or *feature selection* algorithms are dimensionality reduction methods in which the features that are finally employed correspond to original features and

not to a combination. Since the original features often have a clear meaning in the problem domain, those methods are preferable for the interpretation of results.

In this thesis, we will introduce a feature selection method for functional data, based on the work of Berrendero, Cuevas, and Torrecilla, 2016b and Torrecilla and Suárez, 2016, and we will explore its properties, and compare it with other dimensionality reduction methods.

1.3 Structure of the thesis

This thesis is structured as follows.

In chapter 2 we will present the previous work necessary to derive the mathematical properties of the method. We will introduce the optimal classification method, known as the Bayes rule, and show that it often depends on a finite number of points. We will introduce the notion of measures of dependency, and define the ones used in this work. We will also describe Maxima Hunting, a method that share some ideas with the principal method described in this work.

We will introduce the Recursive Maxima Hunting method in chapter 3, with a complete description of its implementation and its most interesting mathematical properties. We will see that this method solves the main flaws of Maxima Hunting.

In chapter 4 we will justify that *Recursive Maxima Hunting* finds the points that appear in the Bayes rule when the mean of the second class is piecewise linear and the noise process correspond to a certain limit of Ornstein-Uhlenbeck processes.

We will show experiments with real and synthetic data in chapter 5 comparing *Recursive Maxima Hunting* with other dimensionality reduction methods. We will also illustrate with small experiments several interesting cases that deserve further study.

The conclusions of this thesis and future work are presented in chapter 6.

Chapter 2

Antecedents and previous work

In this section we will show how for some particular cases the optimal classification rule for functional data depends only on a combination of a finite number of points of the sampled function, thus further justifying our decision for applying a dimensionality reduction method before classifying.

We will also describe some dependency measures between random variables. We will use these dependency measures in the feature selection methods proposed, to find relevant features.

The final section of this chapter will introduce Maxima Hunting. This is a feature selection method that inspired the method proposed in this thesis, Recursive Maxima Hunting.

2.1 Bayes rule

As stated in section 1.1, the objective of binary classification in FDA is to give a *classifier* or *classification rule*, which is a measurable function $g : \mathcal{F} \to \{0, 1\}$ from the feature space $\mathcal{F}$, which is a set of functions, to the set of class labels. The performance of a classifier g is measured by the *classification error* $L = \mathbb{P}(g(X) \neq Y)$, which is the probability of selecting the incorrect class label.

The decision rule that has the lowest classification error is the *Bayes classifier*, or *Bayes rule*

$$g^*(x) = \mathbb{I}_{\{\eta(x) > \frac{1}{2}\}},$$

where $\eta(x) = \mathbb{E}(Y \mid X = x) = \mathbb{P}(Y = 1 \mid X = x)$. Thus, the Bayes rule selects the most probable class given the known data, as we intuitively would expect. The classification error of the Bayes classifier is known as the *Bayes error*. This error is not necessary zero, because in most cases the class of an element f of the feature space $\mathcal{F}$ is not completely determined by f. The Bayes rule is unknown in most classification problems of interest. Therefore, the objective in classification is to provide a reasonable approximation of g^* by induction from the training data.

To compare the performance of two classifiers, one would need to compute their respective classification errors. However, in general the distribution of X and Y is unknown. Therefore the classification error of the classifiers can not be calculated. We should resort instead to calculate the *empirical risk* associated with g

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{g(x_i) \neq y_i\}},$$

over a test sample $\{(x_i, y_i)\}_{i=1}^n$ independent from the training set.

The functional classification problem considered in this study is

$$X(t) = \begin{cases} Z(t), & \text{if } Y = 0, \\ \mu(t) + Z(t), & \text{if } Y = 1, \end{cases} \tag{2.1}$$

where $X(t)$ is the stochastic process generating the observed functions, Y is the random variable corresponding to the class labels, $\mu(t)$ is a deterministic function and $Z(t)$ is a zero-mean stochastic process. We will show that, assuming this form for the classification problem and for particular choices of $\mu(t)$ and $Z(t)$, the Bayes rule depends only on a finite number of points. This means that, if we apply a feature selection step before the classification and the points that appear in the Bayes rule are selected by that method, the best possible classification could still be achieved. If we want to prove that a feature selection method selects these points, we need to calculate the Bayes rule explicitly in those cases. We will calculate it using the concept of Reproducing Kernel Hilbert Spaces (RKHS) and their properties.

2.1.1 Reproducing Kernel Hilbert Spaces (RKHS)

A Reproducing Kernel Hilbert Space $\mathcal{H}$ is a set of functions $f : \mathcal{X} \to \mathbb{R}$ with the following properties (Aronszajn, 1950; Gretton, 2013):

- It is a Hilbert space. This means that $\mathcal{H}$ is a vector space with an inner product $\langle f, g \rangle$, a norm $||f|| = \sqrt{\langle f, f \rangle}$ defined using this inner product, and a distance between elements $d(x, y) = ||x - y||$, defined using the norm. $\mathcal{H}$ must also be a complete metric space with the distance d. That is, every Cauchy sequence in $\mathcal{H}$ converges in $\mathcal{H}$. In summary, a Hilbert space is a Banach space in which the norm is defined using an inner product.

- For every point x in $\mathcal{X}$ we can define a function K_x in $\mathcal{H}$ that has the *reproducing property*; that is, the inner product of every function f in $\mathcal{H}$ with K_x is the evaluation of f in x:

$$\forall x \in \mathcal{X} \quad \exists K_x \in \mathcal{H} \text{ that verifies } f(x) = \langle f, K_x \rangle \quad \forall f \in \mathcal{H}$$

- $\mathcal{H}$ has an associated function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as

$$K(x, y) = \langle K_x, K_y \rangle.$$

  As a result of this point and the reproducing property, for any $x \in \mathcal{X}$ the function K_x is $K(x, \cdot)$.

  The function K is both symmetric and positive definite

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$$

  for every $n \in \mathbb{N}, x_1, \dots x_n \in \mathcal{X}$ and $a_1, \dots, a_n \in \mathbb{R}$. Thus K is a kernel or covariance function.

Given a particular covariance function K, it is possible to derive a unique associated RKHS $\mathcal{H}(K)$ (Aronszajn, 1950) using the Moore-Aronszajn theorem. Given a

symmetric and positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can define the set

$$\mathcal{H}_0(K) = \left\{ f : f(z) = \sum_{i=1}^{n} a_i K(z, x_i) \quad n \in \mathbb{N}, \quad x_i \in \mathcal{X}, \quad a_i \in \mathbb{R}, \quad \sum_{i=1}^{n} a_i K(x_i, x_i) < \infty \right\}.$$

Then for every $f, g \in \mathcal{H}_0(K)$, if $f(z) = \sum_i a_i K(z, x_i)$ and $g(z) = \sum_j b_j K(z, y_j)$, we define the inner product between f and g as

$$\langle f, g \rangle = \sum_{i,j} a_i b_j K(x_i, y_j),$$

and define the norm and distance in $\mathcal{H}_0(K)$ associated with this inner product.

Since any metric space can be completed uniquely, we define $\mathcal{H}(K)$ as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : \mathcal{X} \to \mathbb{R}$ obtained as a pointwise limit of a Cauchy sequence $\{f_i\}$ in $\mathcal{H}_0(K)$. It can be shown that $\mathcal{H}(K)$ satisfies all the properties of an RKHS.

The RKHS $\mathcal{H}(K)$ defined in this way provides a natural inner product that is given by K.

2.1.2 The Bayes rule for equivalent distributions

Given two σ-finite measures μ and ν defined on the same space $\mathcal{X}$, endowed with a σ-algebra $\mathcal{F}$, μ and ν are *equivalent* ($\mu \sim \nu$), or *mutually absolutely continuous* if they have the same zero-measure sets

$$\mu \sim \nu := \forall A \in \mathcal{F} \quad \mu(A) = 0 \iff \nu(A) = 0.$$

The measures μ and ν are *mutually singular* ($\mu \perp \nu$) if there are two sets $A, B \in \mathcal{F}$ such that $A \cap B = \emptyset$, $A \cup B = \mathcal{X}$, and $\mu(B) = \nu(A) = 0$. If μ and ν are probability distributions, μ and ν are *mutually singular* if there exists a set A that has probability 0 with one of the distributions while the same set has probability 1 with the other distribution (the set A^C, the complement of A, plays the role of B in the previous definition). Thus:

$$\mu \perp \nu := \exists A \in \mathcal{F} \quad \mu(A) = 0 \wedge \nu(A) = 1.$$

If μ and ν are Gaussian, then either they are equivalent or they are mutually singular (Feldman, 1958). This result is known as the Hajek-Feldman dichotomy.

If μ and ν are equivalent, there exist a function, usually denoted as $\frac{d\nu}{d\mu}$, with the property

$$\forall A \in \mathcal{F} \quad \nu(A) = \int_A \frac{d\nu}{d\mu} d\mu,$$

which is called the *Radon-Nikodym derivative* of ν with respect to μ. This reflects the fact that $\frac{d\nu}{d\mu}$ is analogous to a derivative in calculus, in the sense that it describes the rate of change of the measure ν with respect to μ.

The Radon-Nikodym derivative provides a way to compute the Bayes rule for binary classification problems in FDA (Baíllo, Cuevas, and Cuesta-Albertos, 2011). This has a special importance in FDA because there is no analog to the Lebesgue

measure in functional spaces (Ferraty and Vieu, 2006). If P_0 and P_1 are the probability distributions of $X \mid Y = 0$ and $X \mid Y = 1$, in the classification problem defined in Equation (2.1), then the Bayes rule is

$$g^*(x) = \mathbb{I}_{\left\{\frac{dP_1}{dP_0}(x) > \frac{1-p}{p}\right\}}, \tag{2.2}$$

where $p = \mathbb{P}(Y = 1)$.

The problem of finding the Bayes rule is now reduced to the problem of finding the Radon-Nikodym derivative of the probability distributions P_0 and P_1. The following theorem by Parzen (Parzen, 1961) can be used to find the Bayes rule for the functional classification problem described in Equation (2.1) assuming that $\frac{dP_1}{dP_0}$ is known and that P_0 and P_1 are Gaussian processes.

Theorem 2.1.1 (Parzen, 1961, Thm 7A)**.** Let P_1 be the distribution of a Gaussian process $\{X(t), t \in \mathcal{T}\}$ with continuous kernel $K(s,t) = \text{Cov}(X(s), X(t))$ and mean function $\mu(t)$. Let P_0 be the distribution of another Gaussian process with the same covariance function and mean 0. Assume that $\mathcal{T}$ is either countable or a separable metric space, and that $\{X(t), t \in \mathcal{T}\}$ is separable. Then:

- $P_0 \sim P_1 \iff \mu \in \mathcal{H}(K)$, where $\mathcal{H}(K)$ is the reproducing kernel Hilbert space associated with K. Note that since the processes involved are Gaussian, this implies that $P_0 \perp P_1 \iff \mu \notin \mathcal{H}(K)$.
- If $P_0 \sim P_1$ the Radon-Nikodym derivative is

$$\frac{dP_1}{dP_0} = \exp\left(\langle X, \mu\rangle_K - \frac{1}{2}||\mu||_K^2\right),$$

  where $\langle \cdot\,, \cdot \rangle_K$ is an operation related with the scalar product in $\mathcal{H}(K)$, as explained below, and $||\cdot||_K$ is the norm in $\mathcal{H}(K)$.

The notation $\langle \cdot\,, \cdot \rangle_K$ is the same as for the inner product in $\mathcal{H}(K)$. In most cases, however, the trajectories x of $X(t)$ are not included, with probability one, in $\mathcal{H}(K)$. Thus, $\langle X, \mu\rangle_K$ is an abuse of notation. Nevertheless, it is possible to give a formal definition of $\langle X, \mu\rangle_K$ (Parzen, 1961) using a linear mapping between the Hilbert space spawned by the family $\{X(t), t \in T\}$ and the space $\mathcal{H}(K)$. The quantity $\langle X, \mu\rangle_K$ has the following properties:

- $\langle X, K(\cdot\,, t)\rangle_K = X(t)$
- $\mathbb{E}(\langle X, h\rangle_K) = \langle \mu, h\rangle_K$
- $\text{Cov}(\langle X, h\rangle_K, \langle X, g\rangle_K) = \langle h, g\rangle_K$

Using Equation (2.2) and Theorem 2.1.1, we can write an expression for the Bayes rule for the classification problem in Equation (2.1) (Berrendero, Cuevas, and Torrecilla, 2017) as $g^*(x) = \mathbb{I}_{\{\eta^*(x) > 0\}}$, where

$$\eta^*(x) = \langle X, \mu\rangle_K - \frac{1}{2}||\mu||_K^2 - \log\left(\frac{1-p}{p}\right). \tag{2.3}$$

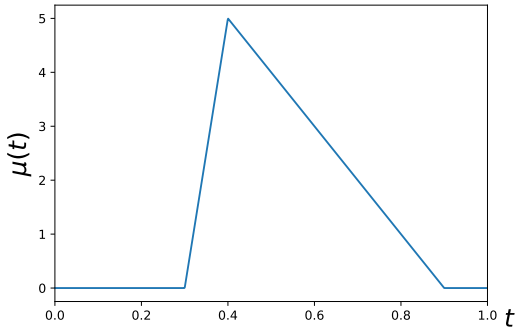

FIGURE 2.1: The mean used in the example. In this case, $s_1 = 0.3$, $s_2 = 0.4$, $s_3 = 0.9$ and $a = \mu(s_2) = 5$.

Example of Bayes rule: Brownian motion with a peak-shaped piecewise linear mean

As an example of how to apply Equation (2.3), let us derive the Bayes rule for a particular example. We consider a classification problem of the family in Equation (2.1), where $Z(t)$ is a standard Brownian process between 0 and 1, with covariance function $K(s,t) = \min(s,t)$, and $\mu(t)$ the piecewise linear function

$$\mu(t) = \begin{cases} 0, & \text{if } t < s_1, \\ a\frac{t-s_1}{s_2-s_1}, & \text{if } s_1 < t < s_2, \\ -a\frac{t-s_2}{s_3-s_2} + a, & \text{if } s_2 < t < s_3, \\ 0, & \text{if } t > s_3, \end{cases}$$

with $a \in \mathbb{R}, a \neq 0$ and $s_1 < s_2 < s_3$. This function correspond to the peak-shaped function shown in Figure 2.1.

(Berlinet and Thomas-Agnan, 2011) For the Brownian kernel $K(s,t) = \min(s,t)$, the associated RKHS $\mathcal{H}(K)$ is

$$\mathcal{H}(K) = \left\{ \begin{array}{l} f(t) \in \mathcal{C}^1[0,1] : f(0) = 0 \quad \wedge \quad f'(t) \in L^2[0,1] \quad \wedge \\ \qquad\qquad f \text{ absolutely continuous } \left(f(t) = f(0) + \int_0^t f'(u)du \right) \end{array} \right\}$$

$$\langle f, g \rangle_K = \int_0^1 f'(t)g'(t)dt$$

Also the operation $\langle X, \mu \rangle_K$ used in Theorem 2.1.1 coincides with the Itô integral

$$\langle X, \mu \rangle_K = \int_0^1 \mu'(t)dX.$$

Now we calculate:

$$\begin{aligned} ||\mu||_K^2 &= \int_0^1 (\mu'(t))^2 dt = \int_{s_1}^{s_2} \left(\frac{a}{s_2 - s_1}\right)^2 dt + \int_{s_2}^{s_3} \left(\frac{a}{s_3 - s_2}\right)^2 dt \\ &= \frac{a^2}{(s_2 - s_1)^2} t \Big|_{s_1}^{s_2} + \frac{a^2}{(s_3 - s_2)^2} t \Big|_{s_2}^{s_3} = a^2 \left(\frac{1}{s_2 - s_1} + \frac{1}{s_3 - s_2}\right), \\ \langle X, \mu \rangle_K &= \int_0^1 \mu'(t) dX = \int_{s_1}^{s_2} \frac{a}{s_2 - s_1} dX - \int_{s_2}^{s_3} \frac{a}{s_3 - s_2} dX \\ &= \frac{a}{s_2 - s_1}(X(s_2) - X(s_1)) - \frac{a}{s_3 - s_2}(X(s_3) - X(s_2)). \end{aligned} \tag{2.4}$$

Using Equation (2.4) and Equation (2.3), the Bayes rule for this problem is $g^*(x) = \mathbb{I}_{\{\eta^*(x) > 0\}}$, where

$$\begin{aligned} \eta^*(x) &= \langle X, \mu \rangle_K - \frac{1}{2}||\mu||_K^2 - \log\left(\frac{1-p}{p}\right) \\ &= \frac{a}{s_2 - s_1}(X(s_2) - X(s_1)) - \frac{a}{s_3 - s_2}(X(s_3) - X(s_2)) - \frac{a^2}{2}\left(\frac{1}{s_2 - s_1} + \frac{1}{s_3 - s_2}\right) \\ &\quad - \log\left(\frac{1-p}{p}\right). \end{aligned}$$

The Bayes rule depends only on $X(s_1)$, $X(s_2)$ and $X(s_3)$. Thus, if a feature selection algorithm selects these three points, it is possible to provide the best possible classification of the trajectories.

Bayes rule under the sparsity assumption

An important case where the Bayes rule depends on a finite number of points of the trajectory in the classification problem given by Equation (2.1) is under the sparsity assumption (Berrendero, Cuevas, and Torrecilla, 2017)

$$\mu(\,\cdot\,) = \sum_{i=1}^{n} a_i K(\,\cdot\,, t_i).$$

where $a_1, \ldots, a_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in \mathcal{T}$.

This assumption is not very restrictive, since the finite combinations of type $\sum_{i=1}^n a_i K(\,\cdot\,, t_i)$ are dense in $\mathcal{H}(K)$; that is, every element of $\mathcal{H}(K)$ can be approximated by these finite combinations. This is because the elements of $\mathcal{H}_0(K)$ verify this assumption, $\mathcal{H}(K)$ is the completion of $\mathcal{H}_0(K)$ and every metric space is dense in its completion. Under this assumption, the only points of X that appear in the Bayes rule are $X(t_1), \ldots, X(t_n)$. The discriminant score $\eta^*(x)$ of a trajectory is

$$\begin{aligned} \eta^*(x) &= \left\langle x, \sum_{i=1}^n a_i K(\,\cdot\,, t_i) \right\rangle_K - \frac{1}{2}\left\| \sum_{i=1}^n a_i K(\,\cdot\,, t_i) \right\|_K - \log\left(\frac{1-p}{p}\right) \\ &= \sum_{i=1}^n a_i \langle x, K(\,\cdot\,, t_i)\rangle_K - \frac{1}{2}\left\langle \sum_{i=1}^n a_i K(\,\cdot\,, t_i), \sum_{j=1}^n a_j K(\,\cdot\,, t_j) \right\rangle_K - \log\left(\frac{1-p}{p}\right) \\ &= \sum_{i=1}^n a_i x(t_i) - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n a_i a_j K(t_i, t_j) - \log\left(\frac{1-p}{p}\right), \end{aligned}$$

where in the last step we have used the reproducing property. This example will be referenced later, as in chapter 4 we will prove that given a particular set of conditions, when the mean verifies this sparsity assumption, our proposed feature selection method selects the points that appear in the Bayes rule.

2.1.3 The mutually singular case: near perfect classification

We have already explained how to obtain the Bayes rule when we have two Gaussian distributions P_0 and P_1 that are equivalent ($P_0 \sim P_1$). As said earlier, if P_0 and P_1 are Gaussian and are not equivalent, then they are mutually singular ($P_0 \perp P_1$). We will now center our attention in that case.

In FDA there are non-trivial classification problems in which it is possible to achieve asymptotic perfect classification (Delaigle and Hall, 2012). This means that for such problems, we can construct simple classifiers based on train data, in which the probability of correctly classifying a function tends to 1 as the size of the training set increases. In that case, basic linear methods for classification become optimal. This phenomenon is called *near-perfect classification*. This is in contrast with the finite-dimensional case, where only pathological examples have this property.

The near-perfect classification phenomenon in binary functional classification has been linked to the relationship between the probability distributions P_0 and P_1 of the intervening classes. In fact, it has been shown that this near-perfect phenomenon appears if and only if P_0 and P_1 are mutually singular (Torrecilla, 2015; Berrendero, Cuevas, and Torrecilla, 2017).

Thus, if P_0 and P_1 are Gaussian, and the mean μ is not in $\mathcal{H}(K)$, by Theorem 2.1.1 $P_0 \perp P_1$ and so we are in a case of near-perfect classification.

2.2 Dependency measures

The objective of feature selection is to identify a subset of variables that captures the information of the original complete set of variables. In the case of classification, those variables can together still make good class predictions, and make the results easier to interpret (Guyon and Elisseeff, 2003). For achieving this, the variables in the set must be *relevant*, that is, the variables must be useful to predict the class. In the other hand, is desirable, to make this set smaller, that the variables are not *redundant*, so that one variable becomes useless once the other variables are known. These objectives are opposing, in the sense that adding relevant variables might increase the redundancy between the selected variables, and removing redundant variables might reduce the predictive capabilities of the selected subset.

There is not a single way of measuring redundancy or relevance. However, there exist some measures of dependency between random variables that could be used to gain some insight into which variables are redundant or relevant. Given one of these measures, we could say that a relevant variable has a strong dependency with the random variable corresponding to the class. In a similar way, two variables are redundant if they have a strong dependency between them.

One example of a dependency measure is the widely used Pearson correlation coefficient, defined for two random variables X and Y as

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

This dependency measure gives values in the interval $[-1, 1]$. Values -1 or 1 are given when the relation between the random variables can be described perfectly by a (non constant) linear equation. A value of 0 implies that there is no linear relationship between the random variables. However, Pearson correlation coefficient measures only linear dependencies between the random variables, and thus does not characterize independence, except when X and Y are jointly normal. Two random variables can have a zero correlation coefficient even if they are dependent.

We will now present some dependency measures that can detect some non linear dependencies and characterize independence, along with their most important properties.

2.2.1 Mutual information

The mutual information between two continuous random variables X and Y is defined (Cover and Thomas, 2012) as

$$I(X,Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy, \tag{2.5}$$

where $p(x)$, $p(y)$ and $p(x,y)$ are the appropriate density functions. If X and Y are discrete random variables, the mutual information is written as

$$I(X,Y) = \sum_{i,j} \mathbb{P}(x_i, y_j) \log \left(\frac{\mathbb{P}(x_i, y_j)}{\mathbb{P}(x_i)\mathbb{P}(y_j)} \right), \tag{2.6}$$

where $\{x_i\}_{i=0}^n$ and $\{y_j\}_{i=0}^n$ are the sets of possible values that X and Y can take respectively.

Some of the properties of Mutual information are:

- $I(X,Y) \geq 0$.
- $I(X,Y) = 0$ if and only if X and Y are independent.
- $I(X,Y) = I(Y,X)$.

The formula of MI in Equation (2.5) is the Kullback-Leibler divergence between the joint distribution and the product of the marginals. MI can also be interpreted in terms of entropy and conditional entropy. The entropy of a discrete random variable X which takes values $x_1 \dots x_n$ is defined as

$$H(X) = -\sum_{i=1}^{n} \mathbb{P}(x_i) \log(\mathbb{P}(x_i)).$$

The entropy measures the average length of the shortest description of the random variable. If the logarithm is base 2, the length is given in bits. It can also be thought as the amount of randomness of the random variable, as more predictable random variables require less bits to describe them (Shannon, 2001).

A related concept is the conditional entropy of a discrete random variable X given another discrete random variable Y, which is defined as

$$H(X \mid Y) = -\sum_{i,j} \mathbb{P}(x_i, y_j) \log \left(\frac{\mathbb{P}(x_i)}{\mathbb{P}(y_j)} \right).$$

This quantity measures the average length of the shortest description of the random variable X if the value of the variable Y is known.

Thus, in the discrete case, we have that

$$I(X,Y) = \sum_{i,j} \mathbb{P}(x_i, y_j) \log \left(\frac{\mathbb{P}(x_i, y_j)}{\mathbb{P}(x_i)\mathbb{P}(y_j)} \right) = H(X) - H(X \mid Y)$$

Thus, the mutual information measures the length of the information about X given by this value of Y. It follows that the value of the mutual information is higher if the two variables have a strong dependency between them.

The definition of MI for continuous variables presented in Equation (2.5) can be obtained as the limit of the discrete mutual information of partitions of the random variables X and Y as these partitions become finer (Cover and Thomas, 2012). Thus, the usual approach for estimating MI with continuous random variables X and Y is to give a partition of the support of the variables into bins of finite size (Butte and Kohane, 2000, Michaels et al., 1998). Then the continuous MI given by Equation (2.5) can be approximated with the discrete version of MI given by Equation (2.6). It has been proved that this approach has systematic errors (Roulston, 1999) independent of the underlying probability distribution. Also this approach requires to provide the number of bins, whose optimal value is not easy to determine. Other approaches for estimating MI replace the bins with more complicated methods such as kernel estimations (Moon, Rajagopalan, and Lall, 1995), or k-nearest neighbor distances (Kraskov, Stögbauer, and Grassberger, 2004) to obtain more accurate results.

We will next show distance covariance, a dependency measure with a simple estimator that does not require parameter estimation and also characterizes independence, thus overcoming the problems of mutual information.

2.2.2 Distance covariance and distance correlation

Distance covariance and distance correlation are recently introduced dependency measures between random vectors (Székely, Rizzo, and Bakirov, 2007). Let X and Y be two random vectors with finite first moments, and let ϕ_X and ϕ_Y be the respective characteristic functions

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$
$$\phi_Y(t) = \mathbb{E}[e^{itY}]$$

Let $\phi_{X,Y}$ be the joint characteristic function. Then, if X and Y take values in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, the distance covariance between them $\mathcal{V}(X,Y)$, or $\mathrm{dCov}(X,Y)$, is the non-negative number defined by

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 w(t,s) dt ds,$$

where $w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$, $|\cdot|_d$ is the euclidean norm in $\mathbb{R}^d$ and $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ is half the surface area of the unit sphere in $\mathbb{R}^d$. The distance correlation $\mathcal{R}(X,Y)$, or $\mathrm{dCor}(X,Y)$, is defined as

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)} & \text{if } \mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y) > 0 \\ 0 & \text{if } \mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y) = 0. \end{cases}$$

We can also use as the dependency measures the squared coefficients $\mathcal{V}^2(X,Y)$ and $\mathcal{R}^2(X,Y)$ directly instead of computing the square root.

The distance covariance has the following properties:

- $\mathcal{V}(X,Y) \geq 0$.
- $\mathcal{V}(X,Y) = 0$ if and only if X and Y are independent.
- $\mathcal{V}(X,Y) = \mathcal{V}(Y,X)$.
- $\mathcal{V}^2(\boldsymbol{a}_1 + b_1\boldsymbol{C}_1X, \boldsymbol{a}_2 + b_2\boldsymbol{C}_2Y) = |b_1b_2|\mathcal{V}^2(Y,X)$ for all constant real-valued vectors $\boldsymbol{a}_1, \boldsymbol{a}_2$, scalars b_1, b_2 and orthonormal matrices $\boldsymbol{C}_1, \boldsymbol{C}_2$.
- If the random vectors (X_1, Y_1) and (X_2, Y_2) are independent then

$$\mathcal{V}(X_1 + X_2, Y_1 + Y_2) \leq \mathcal{V}(X_1, Y_1) + \mathcal{V}(X_2, Y_2).$$

The distance correlation has the following properties:

- $0 \leq \mathcal{R}(X,Y) \leq 1$.
- $\mathcal{R}(X,Y) = 0$ if and only if X and Y are independent.
- If $\mathcal{R}(X,Y) = 1$ then there exists a vector $\boldsymbol{a}$, a nonzero real number b and an orthogonal matrix $\boldsymbol{C}$ such that $Y = \boldsymbol{a} + b\boldsymbol{C}X$.

Distance covariance has an estimator with a simple form. Suppose that we have n observations of X and Y. We denote as X_i the i-th observation of X, and Y_i the i-th observation of Y. If we define $a_{ij} = |X_i - X_j|_p$ and $b_{ij} = |Y_i - Y_j|_q$, the corresponding double centered matrices are defined by $(A_{i,j})_{i,j=1}^n$ and $(B_{i,j})_{i,j=1}^n$

$$A_{i,j} = a_{i,j} - \frac{1}{n}\sum_{l=1}^{n} a_{il} - \frac{1}{n}\sum_{k=1}^{n} a_{kj} + \frac{1}{n^2}\sum_{k=1}^{n} a_{kj},$$
$$B_{i,j} = b_{i,j} - \frac{1}{n}\sum_{l=1}^{n} b_{il} - \frac{1}{n}\sum_{k=1}^{n} b_{kj} + \frac{1}{n^2}\sum_{k=1}^{n} b_{kj}.$$

Then

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{i,j}B_{i,j}$$

is called the squared sample distance covariance, and it is an estimator of $\mathcal{V}^2(X,Y)$. The sample distance correlation $\mathcal{R}_n(X,Y)$ can be obtained as the standardized sample covariance

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y)}, & \text{if } \mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y) > 0, \\ 0, & \text{if } \mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y) = 0. \end{cases}$$

These estimators have the following properties:

- $\mathcal{V}_n^2(X,Y) \geq 0$
- $0 \leq \mathcal{R}_n^2(X,Y) \leq 1$

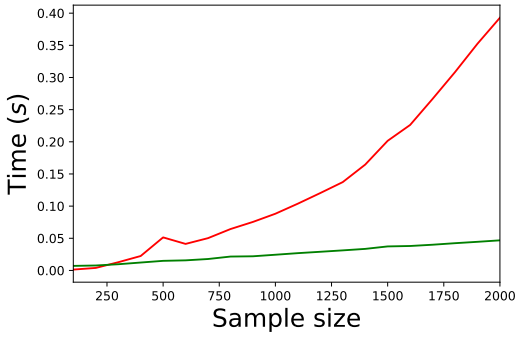

FIGURE 2.2: Distance correlation between two random variables has been estimated for several sample sizes using $\mathcal{V}_n^2(X,Y)$ as the squared distance covariance estimator, in red, and $\Omega_n(X,Y)$ with the faster algorithm, in green. As its complexity is $O(n \log n)$, it outperforms $\mathcal{V}_n^2(X,Y)$, whose complexity is $O(n^2)$, when the number of samples grows.

In a similar way one can define an unbiased estimator $\Omega_n(X,Y)$ of the squared distance covariance $\mathcal{V}^2(X,Y)$. Given the previous definitions of a_{ij} and b_{ij}, we define the U-centered matrices $(\widetilde{A}_{i,j})_{i,j=1}^n$ and $(\widetilde{B}_{i,j})_{i,j=1}^n$

$$\widetilde{A}_{i,j} = a_{i,j} - \frac{1}{n-2}\sum_{l=1}^{n} a_{il} - \frac{1}{n-2}\sum_{k=1}^{n} a_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k=1}^{n} a_{kj},$$
$$\widetilde{B}_{i,j} = b_{i,j} - \frac{1}{n-2}\sum_{l=1}^{n} b_{il} - \frac{1}{n-2}\sum_{k=1}^{n} b_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k=1}^{n} b_{kj}.$$

Then, $\Omega_n(X,Y)$ is defined as

$$\Omega_n(X,Y) = \frac{1}{n(n-3)}\sum_{i,j=1}^{n} \widetilde{A}_{i,j}\widetilde{B}_{i,j}.$$

We can also obtain an estimator of $\mathcal{R}^2(X,Y)$ using $\Omega_n(X,Y)$, as we did with $\mathcal{V}_n^2(X,Y)$. $\Omega_n(X,Y)$ does not verify that $\Omega_n(X,Y) \geq 0$, because sometimes could take negative values near 0. In our case, this is not a problem because we do not rely on this property. The main advantage of using $\Omega_n(X,Y)$ over $\mathcal{V}_n^2(X,Y)$ is that there is an algorithm that can compute $\Omega_n(X,Y)$ for random variables with $O(n \log n)$ complexity (Huo and Székely, 2016). Since the estimator formulas explained above have complexity $O(n^2)$, this improvement is significant, specially for larger samples, as shown in Figure 2.2. We will use $\Omega_n(X,Y)$, computed with the faster algorithm, as our estimator for $\mathcal{V}^2(X,Y)$.

Equivalent expression for distance covariance

There is a more convenient expression for calculating the distance covariance between the random variable corresponding with a feature and the class label, using the following theorem (Berrendero, Cuevas, and Torrecilla, 2016b).

Theorem 2.2.1 (Alternative expression for distance covariance)**.** In the setting of the binary functional classification problems described in section 1.1 the distance covariance between a point of X and the class labels Y, $\mathcal{V}^2(X(t), Y)$, can be alternatively calculated as

$$\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2} \right],$$

where $p = \mathbb{P}(Y = 1)$, $I_{ij}(t) = \mathbb{E}(|X(t) - X'(t)|| \; Y = i, Y' = j)$ and (X', Y') is an independent copy of (X, Y).

Using Theorem 2.2.1 we can derive a explicit formula for the distance covariance in the setting of binary functional classification problems, that will be useful in our theoretical derivations. The proof of this result is in Appendix B.

Corollary 2.2.2 (Explicit formula for distance covariance)**.** Let $\mathcal{V}^2$ be the distance covariance function. Under the model given by Equation (1.1) $\mathcal{V}^2(X(t), Y)$ has the following expression:

$$\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[\frac{2\sigma(t)}{\sqrt{\pi}} \left(e^{-\frac{\mu(t)^2}{4\sigma(t)^2}} - 1 \right) + \mu(t) \left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)} \right) - 1 \right) \right]$$

where $p = \mathbb{P}(Y = 1)$, cdf is the cumulative distribution function of a standard normal random variable, and $\sigma(t)$ is the standard deviation of the noise process $Z(t)$ at point t.

Also, the first and second derivatives are

$$\begin{aligned}
&\frac{d}{dt}\mathcal{V}^2(X(t), Y) \\
&\quad = 4p^2(1-p)^2 \left[\frac{2\sigma'(t)}{\sqrt{\pi}} \left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1 \right) + \mu'(t) \left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)} \right) - 1 \right) \right], \\
&\frac{d^2}{dt^2}\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[\frac{2\sigma''(t)}{\sqrt{\pi}} \left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1 \right) \right. \\
&\qquad\qquad + \frac{1}{\sqrt{\pi}} e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} \left(\frac{(\mu(t)\sigma'(t) - \sigma(t)\mu'(t))^2}{\sigma^3(t)} \right) \\
&\qquad\qquad \left. + \mu''(t) \left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)} \right) - 1 \right) \right],
\end{aligned}$$

provided that all the derivatives appearing in the expression exist.

2.3 Maxima Hunting

In Berrendero, Cuevas, and Torrecilla, 2016b a new method of variable selection with functional data, called Maxima Hunting (MH) was proposed. This method is purely functional data method and takes into account the particular structure of this kind of data.

Maxima Hunting first computes a measure of dependency between each variable and the class. This provides a function that measures a kind of *relevance* of each

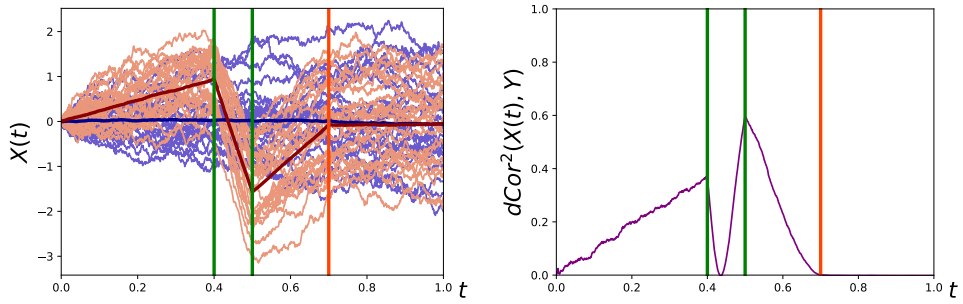

FIGURE 2.3: At the left are shown the sample trajectories to with MH is applied. The right side shows the relevance function whose local maxima will be selected, using distance correlation as a dependency measure. The points selected by MH are shown in both figures with green vertical lines. The point shown with a vertical orange line is a relevant point that appears in the Bayes rule but was not selected by MH.

variable for predicting the class. Then, the method selects the variables that correspond to the local maxima of that function. By selecting only the local maxima, the method guarantees that the selected maxima are more relevant than the close variables, while eliminating the *redundancy* not selecting the nearby variables, which are highly correlated if the sampled functions are smooth.

Based on their experiments, the authors recommended using distance correlation as the measure of dependency. An example using distance correlation is shown on Figure 2.3.

MH has some interesting properties:

- As said above, it selects *relevant* variables and does not select nearby *redundant* variables, as these are not local maxima.
- If the measure used allows us to quantify dependence between random vectors, MH could be applied to multiple classification and to regression problems. Also, it could be applied when the input functions have more than one dimension.
- The estimator of the distance covariance converges uniformly to the real value of the distance covariance (Berrendero, Cuevas, and Torrecilla, 2016b). Thus, the local maxima selected by MH also converge to the real local maxima when distance covariance is used as the dependency measure.

However, the main limitation of MH arises in cases where some variables are not relevant by themselves, but are relevant once some other variables are also selected. As these variables are not local maxima of the relevance function, they will not be selected by MH. One example of this situation is shown on Figure 2.3.

Also, Maxima Hunting has a technical difficulty that must be taken into account. Estimating the local maxima of the relevance function could be a very difficult task, especially if this function is not smooth enough or vary abruptly. In that case, many points could be interpreted as local maxima, and so MH will select too many variables, as shown in Figure 2.4. A possible way to correct this problem is computing the local maxima after the function has been smoothed. If the function

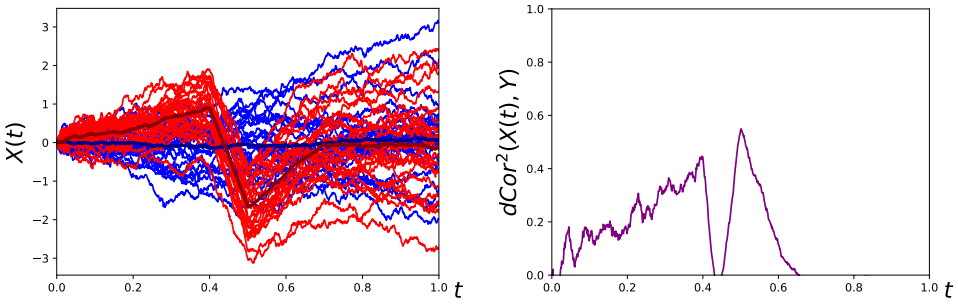

FIGURE 2.4: If the relevance function (shown at the right) is not smooth, MH could mistake for local maxima points that are not true local maxima. This is often the case when the sample size is small.

is smoothed too much, MH could miss perfectly valid local maxima, but if it is not smooth enough, MH could mistake ordinary points for local maxima.

In the next chapter, we will introduce Recursive Maxima Hunting (RMH), a method inspired by MH that overcomes the flaws presented above.

Chapter 3

Recursive Maxima Hunting (RMH)

Recursive Maxima Hunting (RMH) is a filter variable selection method for functional classification problems which is based on Maxima Hunting (Torrecilla and Suárez, 2016). The goal of Maxima Hunting (MH) is to select those features whose dependency with the class labels reach a local maximum (Berrendero, Cuevas, and Torrecilla, 2016b). This dependency can be computed using some measure of dependency, such as mutual information, or distance correlation. The main drawback of Maxima Hunting is that it can not identify features that are not important by themselves but become important after other relevant features have been chosen. Recursive Maxima Hunting (RMH) overcomes this problem by removing at each stage of the process the information on the class provided of the feature that has been selected, thus revealing new relevant features.

RMH can be seen as an iterative version of MH. In the first step of RMH, the objective is to select the feature that best discriminates between the classes. Thus, RMH begins by selecting the point in the trajectories whose dependency with the class is the highest. The trajectories are then corrected by subtracting the influence of the selected point. This process uncovers other features that were not relevant on their own, but are relevant in combination with the selected variables. For the corrected trajectories, these points should have a greater dependency with the class labels and therefore be selected. This process can be repeated until no point in the trajectories has a high dependency with the class. The set of selected features then would be composed by points that are relevant when taken together, but need not to be relevant on their own. Besides being able to identify such variables, RMH requires finding global rather than local maxima, which is done numerically.

The method was introduced in Torrecilla and Suárez, 2016, in the context of binary classification, assuming that the trajectories from each class are a Gaussian noise processes with a different mean, assuming homoscedasticity. In this thesis, we carry out an extensive theoretical analysis of the method and provide further empirical evidence of its effectiveness.

3.1 Description of RMH

In the classification problem described in section 1.1 one has trajectories that belong to two different classes. Thus, the class label is a realization of the dichotomic random variable $Y \in \{0, 1\}$. the trajectories are assumed to be realizations of some stochastic process $X(t)$ dependent on the class

$$X(t) = \begin{cases} Z(t) & \text{if the class is } 0 \ (Y = 0) \\ \mu(t) + Z(t) & \text{if the class is } 1 \ (Y = 1). \end{cases} \tag{3.1}$$

where $\mu(t)$ is a deterministic function and $Z(t)$ is a zero-mean stochastic noise process that is the same for the two populations (homoscedasticity assumption).

As said earlier, RMH is an iterative algorithm, in which the trajectories are modified at each iteration. Thus, the stochastic process that describes those trajectories also changes. We will say that $X^{[i]}$ is that stochastic process after i corrections have been applied, and that t_i is the i-th point selected. The noise process corresponding to the ith correction is $Z^{[i]}$. Similarly, the mean of $X^{[i]}$ when the class is $Y = 1$ is denoted as $\mu^{[i]}$.

Before the first iteration we have

$$\begin{aligned} X^{[0]}(t) &= X(t) \\ Z^{[0]}(t) &= Z(t) \\ \mu^{[0]}(t) &= \mu(t). \end{aligned}$$

At each iteration we select the point t_i using the rule

$$t_i = \underset{t\in[0,1]}{\operatorname{argmax}}\{\text{dependency\_measure}(X^{[i-1]}(t), Y)\},$$

were $\text{dependency\_measure}$ is a dependency measure between random variables.

The process after i iterations have been performed is

$$X^{[i]}(t) = X^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right], \tag{3.2}$$

where $\mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right]$ is the correction applied.

In RMH we will assume that $Z(t)$ is a zero-mean Gaussian process characterized by the kernel (covariance function) $K(s,t)$. In that case, there is an explicit formula for the correction

$$\mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right] = \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)} X^{[i-1]}(t_{i+1}), \tag{3.3}$$

where $K^{[i-1]}(t,s)$ is the covariance function associated with $Z^{[i-1]}$, which is also a zero-mean Gaussian process. By definition, $K^{[0]}(s,t) = K(s,t)$. In practice, this assumption is not too restrictive because it is often possible to approximate the noise process in functional classification problems by a Gaussian process with an appropriate kernel. Furthermore, RMH can be extended to cases where $Z(t)$ is not Gaussian, providing that the appropriate correction is applied at each iteration.

3.1.1 An illustrative example: piecewise linear mean

Before describing the method in detail, we will illustrate the application of RMH in a simple binary classification task. The problem consists in discriminating between standard Brownian trajectories and Brownian trajectories with a piecewise linear mean.

The plots in Figure 3.1 display the evolution of the algorithm. Each row corresponds to an iteration of RMH. The plots on the first column display the trajectories, which are modified at each iteration by applying the correction Equation (3.2). The original trajectories are shown in the first row of this column. Subsequent plots correspond to corrected trajectories. The figures on the second column show the dependence (measured using the squared distance correlation) between the vector of trajectories at each point and the class vector.

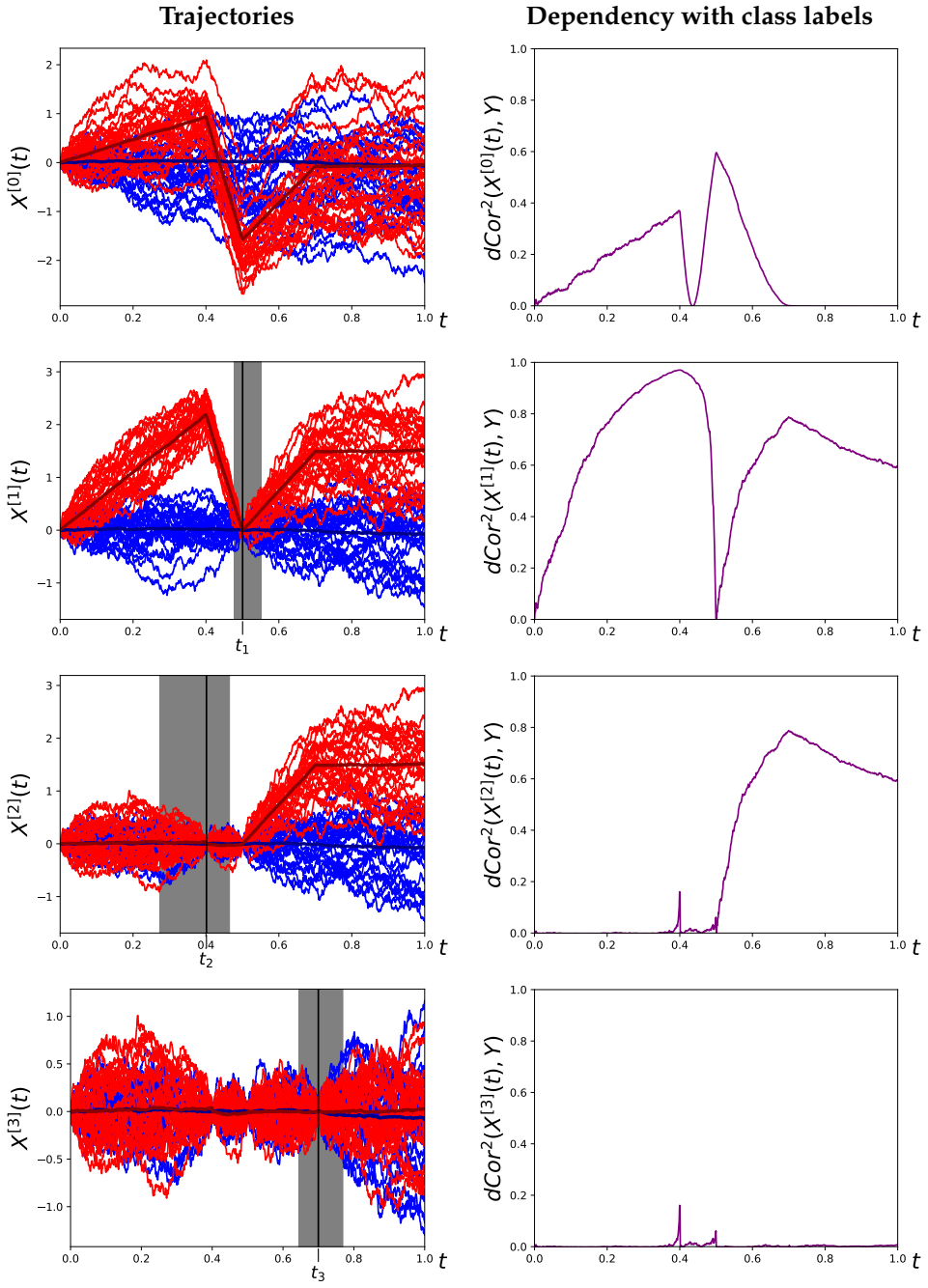

FIGURE 3.1: Example of the execution of RMH with Brownian trajectories corresponding with two classes with different means.

The plot on the first row and second column displays the original trajectories, corresponding to $X^{[0]}(t) = X(t)$. Class 0 trajectories are shown in blue and class 1 trajectories in red. The mean of each class is marked with a darker and wider line. Class 0 trajectories have zero mean, and class 1 trajectories have a nonzero piecewise linear mean $\mu^{[0]}(t) = \mu(t)$. The noise process $Z(t)$ for both kinds is a standard Brownian motion, which is characterized by the covariance function $K(t, s) = \min(t, s)$. The noise process assumed by RMH to apply the corrections is also standard Brownian motion. The plot on the second column in the first row is the squared distance correlation between the points in the trajectory and the class label. Note that the two maxima of this dependency measure coincide with the maxima of the differences between the two means $|\mu(t) - 0|$. Furthermore, the derivative of $\mu(t)$ at these maxima is discontinuous.

On the second row, the point t_1 corresponding with the largest maximum has been selected. This point is marked in the plot with a vertical black line. The grey stripes correspond to points in the neighborhood of the selected one, and are strongly correlated with it. Because of their redundancy, they are discarded for further consideration by the algorithm. The conditional expectation given the selected point, $\mathbb{E}\left[Z^{[0]} \mid Z^{[0]}(t_1) = X^{[0]}(t_1)\right]$, corresponding to a standard Brownian process with zero mean has been subtracted to the trajectories. Since the original noise process is Brownian motion, the noise process for the corrected trajectories is, to the left of the selected point, a Brownian bridge between zero and the selected point, and, to the right of the selected point, a Brownian process beginning at the this point. On the second column one can see that the correlation between the points and the class has two maxima, the maximum that was not selected in the previous step and a new one. The new maximum corresponds also to a point in which the derivative of $\mu(t)$ is discontinuous. This maximum appear only after the first point is selected. Therefore,this point would not have been selected in Maxima Hunting. Between these two maxima, RMH selects the one with the highest correlation.

The third row is similar to the second row. We can see that the subtraction of the conditional expectation has not modified the interval at the right of the first selected point. As will be shown in subsection 3.3.3, this only happens when the assumed noise process is Markovian. The right figure shows that, except for spurious fluctuations, related to the finite size of the samples and numerical amplification of the errors, there is only one maximum in the correlations with the class.

On the fourth row one can see that there are no more relevant point because no significant correlations remain. Thus, the method returns the three selected points. Here, it is important to note that, although there is a small peak caused by numerical errors, it is not selected because it is highly correlated with a previously selected point.

3.2 Implementation of RMH

The pseudocode of the implementation of RMH used in this work is shown in Algorithm 1. The algorithm receives as input a matrix `X`. The rows of `X` correspond to trajectories. The columns correspond to values of the trajectories at different values of t. These values are the features considered for selection. The algorithm receives also the vector `Y` consisting of the class labels of the trajectories.

Another input parameter of RMH is `dependency_measure`. This is a function that quantifies the level of dependency between two random variables. This function is used to measure the dependence between each feature and the class, and

Algorithm 1 Recursive Maxima Hunting

function RMH()̇

▷ Input:
▷ `X`: Set of trajectories
▷ `Y`: Vector with a class label per trajectory
▷ `dependency_measure`: Measure of dependency between two vectors
▷ `correction`: Type of correction applied
▷ `min_relevance`: Threshold for considering a point relevant
▷ `min_redundancy`: Threshold for considering a point redundant with the selected one
▷

▷ Output:
▷ `points`: List with the points selected, in selection order

`mask` $\leftarrow$ `{}` ▷ Set of discarded points
`points` $\leftarrow$ `[]` ▷ List of selected points

▷ Select the best point
$\texttt{t}_{\texttt{max}} \leftarrow \underset{\texttt{t}\in[0,1]}{\text{argmax}}\{\texttt{dependency\_measure}(\texttt{X}(\texttt{t}), \texttt{Y})\}$
do
▷ Discard nearby points strongly correlated with $\texttt{X}(\texttt{t}_{\texttt{max}})$
$\texttt{mask} \leftarrow \texttt{mask} \cup \sup\{[a,b] : \texttt{t}_{\texttt{max}} \in [a,b] \wedge \forall r \in (a,b)$ $\texttt{dependency\_measure}(\texttt{X}(r), \texttt{X}(\texttt{t}_{\texttt{max}})) > \texttt{min\_redundancy}\}$

▷ Apply correction
$\texttt{X} \leftarrow \texttt{apply}(\texttt{correction}, \texttt{X}, \texttt{t}_{\texttt{max}})$

▷ Add the selected point to result list
$\texttt{points} \leftarrow \texttt{points} + [\texttt{t}_{\texttt{max}}]$

▷ Update correction
$\texttt{correction} \leftarrow \texttt{update}(\texttt{correction}, \texttt{t}_{\texttt{max}})$

▷ Select the next best point
$\texttt{t}_{\texttt{max}} \leftarrow \underset{\texttt{t}\in[0,1]\setminus\texttt{mask}}{\text{argmax}}\{\texttt{dependency\_measure}(\texttt{X}(\texttt{t}), \texttt{Y})\}$
while $\texttt{dependency\_measure}(\texttt{X}(\texttt{t}_{\texttt{max}}), \texttt{Y}) > \texttt{min\_relevance}$

▷ Return when no more relevant variables can be identified
return `points`
end function

the dependence between the selected feature and the nearby ones. In our work, we have used the squared distance correlation, which takes values in the interval $[0, 1]$. Other dependency measures can be used, such as the Pearson correlation coefficient, which measures only linear dependencies, mutual information and distance covariance. However, these last two measures are unbounded. In consequence, a more elaborate approach would be required to quantify the relevance and redundancy of the features, rather than simply using the threshold parameters `min_relevance` and `min_redundancy`.

The parameters `min_relevance` and `min_redundancy` are used for thresholding. They lie in the range of values that the `dependency_measure` can take. The parameter `min_relevance` is used to determine when the remaining features are irrelevant, so that the search for relevant points can be halted. If the dependency (as measured by `dependency_measure`) between the selected feature and the class is not above `min_relevance` the algorithm finishes and returns the features selected up to that iteration. The parameter `min_redundancy` is used to discard points that are highly correlated with one of the selected points. The algorithm finds the biggest connected space of points that includes the point selected and whose dependency with the point selected is greater than `min_redundancy`. These points are not considered for selection in subsequent iterations. If `dependency_measure` takes values in, for example, the interval $[0, 1]$ then `min_relevance` should be close to 0, so that all sufficiently relevant features are selected. The value of `min_redundancy` should be close to 1, so that points whose dependency with a previously selected point is low are not discarded.

The `correction` parameter is the type of correction that will be applied. When the instruction $\texttt{apply}(\texttt{correction}, \texttt{X}, \texttt{t}_{\texttt{max}})$ is performed, the matrix of trajectories `X` is modified to discard the information conveyed by the selected point. This modification allows that, in subsequent iterations, new maxima in the dependency curve appear. These maxima correspond to features that are not important on their own but are important together with the points selected. Finally, the function `update` returns the appropriately modified correction function.

In this work, two types of corrections have been applied: the *Gaussian Process correction* (or *GP correction*), and the *Uniform-Brownian correction*. Here is a brief description of the two:

- The *Gaussian Process correction* (or *GP correction*) is a correction that assumes that the noise process is a zero-mean Gaussian process Z, with $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z(s)Z(t)] = K(s, t)$. We will show that the process that results from applying the correction Equation (3.3) is also Gaussian. In consequence, the updated correction is also a *Gaussian Process correction*, with zero mean and a different kernel.

- The first iteration of of RMH with a *Uniform-Brownian correction* consists in subtraction from the original process the value of the process at the point for which the dependence with the class label is maximal

$$X^{[1]} = X^{[0]} - X^{[0]}(t_{max}).$$

  In the next iteration, the correction assumes that $Z^{[1]}$ is an isotropic standard Brownian process, emanating from t_{max}. Since we assume that t is a one-dimensional continuous parameter, this corresponds to a Brownian process that has its origin at t_{max} and whose trajectories evolve to the left and to the right of this point.

The Uniform-Brownian correction can be seen as an special case of a GP correction where the noise process verifies

$$\mathbb{E}[Z^{[0]} \mid Z^{[0]}(t_{max}) = z] = z$$

and the noise process after the first correction, $Z^{[1]}(t)$, is an isotropic standard Brownian process. In chapter 4 we will see that, in fact, the Uniform-Brownian correction can be seen as an Ornstein-Uhlenbeck Gaussian process (Uhlenbeck and Ornstein, 1930) in the limit that the lengthscale of the process approaches infinity.

The are some differences between the implementation of RMH presented in Algorithm 1 and the one in Torrecilla and Suárez, 2016 that are worth commenting upon. In the original implementation, an initial selection is made only if at least one location in the trajectory sufficiently relevant. Otherwise the algorithm would return an empty list of selected points In the current implementation, an initial point is always selected. This is probably advantageous because, once the first point has been selected, applying the correction may reveal new points that are relevant only after the first point has been selected, as shown in Figure 3.2.

Also, in the original algorithm, when a new point is selected, the interval considered is divided into two subintervals. Then, RMH could be applied to each of these subinterval independently. However, this can be done only if the assumed noise process is Markovian, because, as illustrated in subsection 3.3.3, in that case the corrections on an interval do not affect the other intervals. If the process noise process is not Markovian, the effect of the correction extends beyond the subinterval considered and affects the whole process.

3.3 Analysis of RMH

In this section, we will carry out a detailed analysis of RMH. We will prove that, assuming that the underlying process is Gaussian, the corrections preserve the form of the functional classification problem at each iteration. In consequence, RMH can be implemented as a recursive process. We will also show that there is an efficient way of computing the corrections. Finally, we will analyze the consequences of assuming that the underlying Gaussian process is Markovian.

3.3.1 Recursive implementation of RMH

First, We show that RMH can be implemented recursively, because the classification problem after correction Equation (3.3) is applied verifies Equation (3.1) and the updated noise process is Gaussian as well.

Let $X^{[i]}$ be the process after i corrections have been applied

$$X^{[i]}(t) = X^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right], \quad i = 1, 2 \dots$$

and that the original noise process $Z^{[0]}(t) = Z(t)$ is a zero-mean Gaussian process with covariance function $K(t, s)$. The following theorem states that if the original noise process is a zero-mean Gaussian process, then the corrected process is also a zero-mean Gaussian process. Thus is justified that the updated *GP correction* is also a *GP correction* albeit with a different kernel.

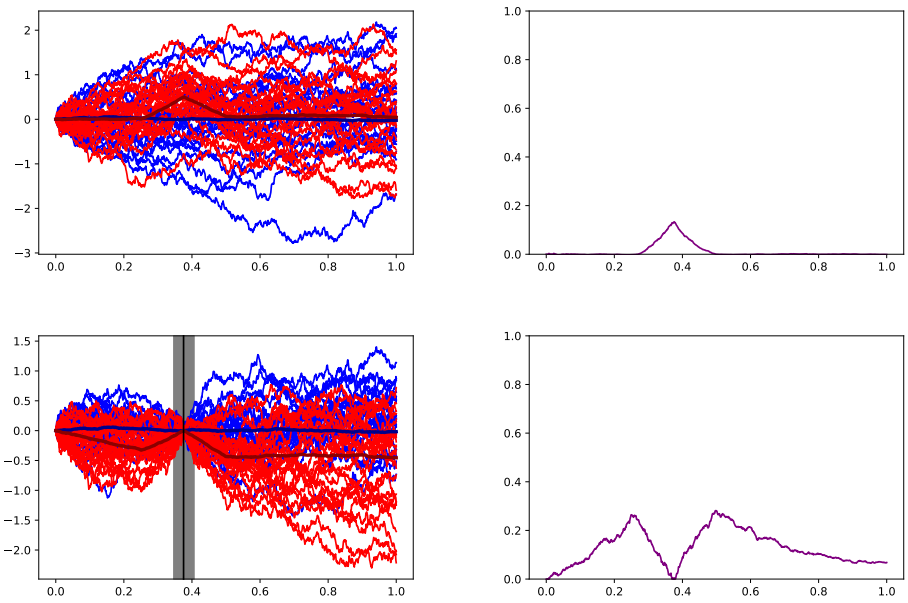

FIGURE 3.2: Unmasking of relevant points: if `min_relevance` were 0.2 the first point would not have been selected. However, selecting it reveals new relevant points for which the dependency with the class labels is above 0.2.

Theorem 3.3.1. If $Z(t)$ is a zero-mean Gaussian process and $X(t)$ verifies Equation (3.1), $X^{[i]}(t)$ for $i \geq 0$ is also of the form specified by Equation (3.1). That is,

$$X^{[i]}(t) = \begin{cases} Z^{[i]}(t) & \text{if the class is } 0 \\ \mu^{[i]}(t) + Z^{[i]}(t) & \text{if the class is } 1, \end{cases}$$

where $Z^{[i]}(t)$ is a Gaussian zero-mean noise process and $\mu^{[i]}(t)$ is a deterministic function, with $X^{[0]}(t) = X(t)$ and $Z^{[0]}(t) = Z(t)$.

Theorem 3.3.2 (Correction formula)**.** For $i \geq 1$, the i-th correction is

$$\mathbb{E}[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)] = \frac{K^{[i-1]}(t, t_i)}{K^{[i-1]}(t_i, t_i)} X^{[i-1]}(t_i),$$

where $K^{[i-1]}(t, s)$ is the covariance function of $Z^{[i-1]}$ and $K^{[0]}(t, s) = K(t, s)$.

Theorem 3.3.3 (Mean and noise formula). For $i \geq 1$, the expression of $Z^{[i]}(t)$ and $\mu^{[i]}(t)$ in Theorem 3.3.1 is

$$\begin{aligned}\mu^{[i]}(t) &= \mu^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = \mu^{[i-1]}(t_i)\right] \\ &= \mu^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}\mu^{[i-1]}(t_i) \\ Z^{[i]}(t) &= Z^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right] \\ &= Z^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}Z^{[i-1]}(t_i).\end{aligned}$$

The proof of the three theorems above is in Appendix B.

3.3.2 An efficient way of computing the corrections

Theorem 3.3.2 states that the i-th correction applied can be expressed as

$$\mathbb{E}[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)] = \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}X^{[i-1]}(t_i),.$$

The above formula has the drawback that one needs to compute all previous covariance functions $K^{[i-1]}, K^{[i-2]}, \ldots, K^{[0]}$. If RMH is implemented directly using this expression, the time to compute one correction increases on each iteration. To make a more efficient implementation of RMH we will provide a way to compute the corrections using only the original covariance function. We will also provide a formula for computing several corrections at once (see Figure 3.3), which is useful for showing the properties of the corrected trajectories. Since the noise process is the same for both classes, in this subsection we will assume for simplicity that the class is 0, that is, $X^{[i]}(t) = Z^{[i]}(t)$. Given that assumption, the correction is

$$\mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right].$$

For class 0 trajectories, it can be rewritten as

$$\mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right].$$

We will also make use of Theorem 3.3.3 for computing the noise process on each iteration as

$$Z^{[i]}(t) = Z^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right]$$

Theorem 3.3.4 (Correction expression at step n given the initial Gaussian process). An alternative, non-recursive expression for $Z^{[n]}$ is

$$Z^{[n]}(t) = \left[Z^{[n-1]}(t) \mid Z^{[n-1]}(t_n) = 0\right] = \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0, \ldots Z^{[0]}(t_n) = 0\right].$$

Thus:

$$[Z^{[n]}(t) \mid Z^{[n]}(s)] = \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0, \ldots Z^{[0]}(t_n) = 0, Z^{[0]}(s) = Z^{[n]}(s)\right].$$

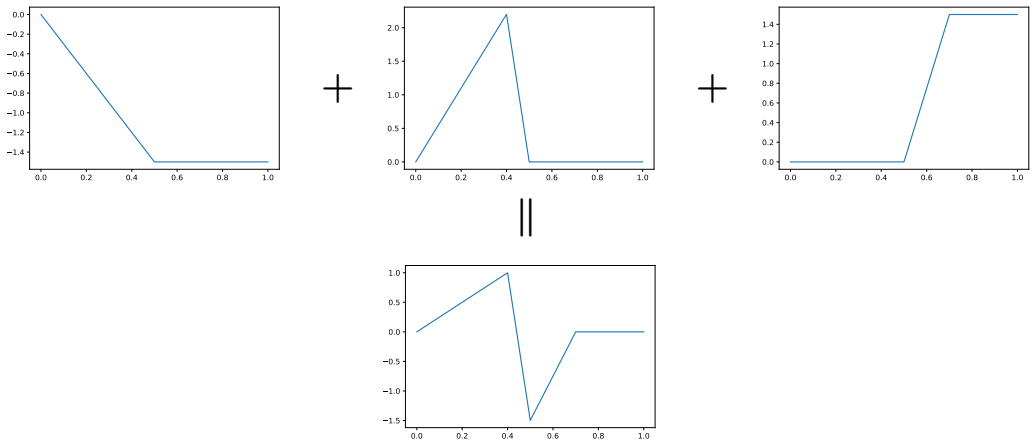

FIGURE 3.3: Theorem 3.3.5 provides a formula to compute several corrections at once, given the initial assumed underlying process and the selected points. Here there are on the first row the means of the corrections applied in the example for the trajectories of the second class, and on the second row the sum of those means (that is the same as the mean of the second class, in this particular example).

Theorem 3.3.5 (Apply several corrections at once)**.** The following equality holds:

$$\mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1), \dots Z^{[0]}(t_n)\right] = \sum_{i=0}^{n-1} \mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1})\right].$$

Theorem 3.3.5 provides a formula for computing several corrections at once.

To prove Theorem 3.3.4 and Theorem 3.3.5, we need to first prove the following lemma.

Lemma 3.3.6 (Alternative expression of a GP correction)**.** Let Z be a Gaussian stochastic process that has mean 0 and covariance function K. Then the process

$$Z(t) - \mathbb{E}\left[Z(t) \mid Z(t_1) \dots Z(t_n)\right]$$

has the same distribution as the process

$$\left[Z(t) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right].$$

The proofs of the lemma and the above theorems are in Appendix B.

3.3.3 RMH with a Markovian Gaussian process

If the trajectories considered are of the form Equation (3.1) with $Z(t)$ a zero mean Gaussian process that is Markovian, the corrected trajectories are of the same form. According to Theorem 3.3.4, the corrected noise processes $Z^{[i]}(t)$, $i \geq 0$ are the result of conditioning the original noise process. Therefore, they are also Markovian.

As mentioned in section 3.1, when the underlying Gaussian process is Markovian, after the correction associated to t_1, the first relevant point, has been applied

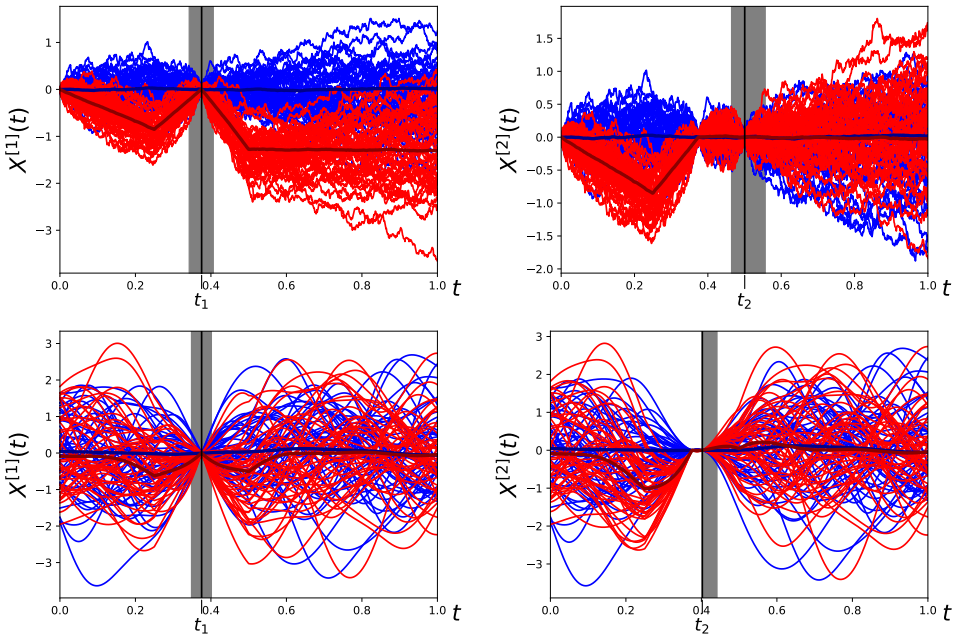

FIGURE 3.4: The first row shows an iteration of RMH using the Brownian correction, which has the Markov property. It is apparent that the correction over the right interval did not change the left interval. The second row shows a RBF correction, which is not Markovian. In this case the correction affects the two intervals. In both cases the assumed noise process is the real one, with zero mean.

it is possible to consider the points to the right of t_1 and to the right of t_1 separately: Because of the Markovian property, $X^{[1]}(t)$ with $t \in [0, t_1]$ and $X^{[1]}(t')$ with $t' \in [t_1, 1]$ are independent. If, in a posterior step, RMH selects a point in one of these subintervals, the corresponding correction will not alter the trajectories in the other subinterval. Since the corrected process is also Markovian, this segmentation also applies for subsequent iterations. Figure 3.4 provides an illustration of this property.

The following theorem states that if the point selected is t, then the points at $[0, t)$ and the points at $(t, 1]$ are independent, because their covariance is zero. Since the noise process is Gaussian, all dependencies are linear, and having a zero covariance characterizes independence. Therefore, the corrections applied in one of the two subintervals do not affect the values of the trajectory at points in the other subinterval. The proof of this theorem is in Appendix B.

Theorem 3.3.7. Let Z be a Gaussian Markov process with covariance function K. Then, for all s, t, u with $s < t < u$ the process $Z - \mathbb{E}[Z \mid Z(t)]$, whose covariance function is K_1, verifies $K_1(s, u) = 0$.

3.4 The GP correction and interpolation of the mean

The corrections of the trajectories carried out at each iteration can be seen as performing an interpolation of the unknown class 1 mean. To illustrate this point, we recall that, if the underlying Gaussian process is Z the correction after $t_1, \ldots, t_n$ are

selected is $\mathbb{E}[Z \mid Z(t_0), \dots Z(t_n)]$ (see Theorem 3.3.5). If this correction is applied to the mean $\mu^{[i]}$ in Theorem 3.3.3, one obtains

$$\mu^{[n]}(t) = \mu^{[0]}(t) - \mathbb{E}\left[Z^{[0]} \mid Z^{[0]}(t_1) = \mu^{[0]}(t_1), \dots, Z^{[0]}(t_n) = \mu^{[0]}(t_n)\right].$$

If the mean $\mu^{[0]}(t)$ and $\mathbb{E}\left[Z^{[0]} \mid Z^{[0]}(t_1) = \mu^{[0]}(t_1), \dots, Z^{[0]}(t_n) = \mu^{[0]}(t_n)\right]$ are close to each other, then $\mu^{[n]}(t)$ will be close to 0. Therefore, after n iterations, the corrected process is mostly noise, and so the dependency between each point of the process and the class will be close to 0 and RMH halts.

This reasoning affords a novel perspective on the corrections: The points selected by RMH, $t_1, \dots, t_n$, are such that

$$f(t) = \mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = \mu^{[0]}(t_1), \dots, Z^{[0]}(t_n) = \mu^{[0]}(t_n)\right]$$

is close to $\mu(t)$, the original mean for class 1 trajectories. This function $f(t)$ interpolates $\mu(t)$ between the points that have been selected by RMH.

The form of the interpolation obtained depends on the particular kernel considered (see Appendix E for a list of different types of kernels and their parameters). This is illustrated in Figure 3.5 and Figure 3.6, which display interpolations of the sine function with different kernels (e.g. "Brownian interpolation", "RBF interpolation", etc.).

As discussed earlier, the choice of kernel determines the shape of the interpolation. For example, both exponential and RBF kernels correspond to stationary Gaussian processes with reversion to the mean. In a Gaussian process that is mean-reverting, trajectories that are initially forced to take values away from their mean tend to approach this mean in a characteristic time that is related to the *lengthscale* parameter of the kernel of the precess. This phenomenon is illustrated in Figure 3.6: The larger the lengthscale the longer it takes for the interpolation to approach zero (the mean of the noise process). When the lengthscale tends to infinity, the interpolation tends to a piecewise linear function.

3.4.1 Interpolation using the Brownian process

Let us analyze the interpolation when the noise process is a Brownian process. From Figure 3.5, it is apparent that the Brownian interpolation is a piecewise linear function. The next result will provide a formal statement of this fact.

Theorem 3.4.1 (Brownian process and linear interpolation)**.** If Z is a Brownian process with mean 0 then $\mathbb{E}[Z \mid Z(t_1) = \mu_1, \dots Z(t_n) = \mu_n]$, assuming $t_i < t_j$ if $i < j$ is the function $f(t)$

(i) $f(t_i) = \mu_i$.

(ii) If $t \in (t_i, t_{i+1})$ then $f(t)$ is the linear interpolation between μ_i and μ_{i+1}

$$f(t) = \mu_i \left(1 - \frac{t - t_i}{t_{i+1} - t_i}\right) + \mu_{i+1} \left(\frac{t - t_i}{t_{i+1} - t_i}\right).$$

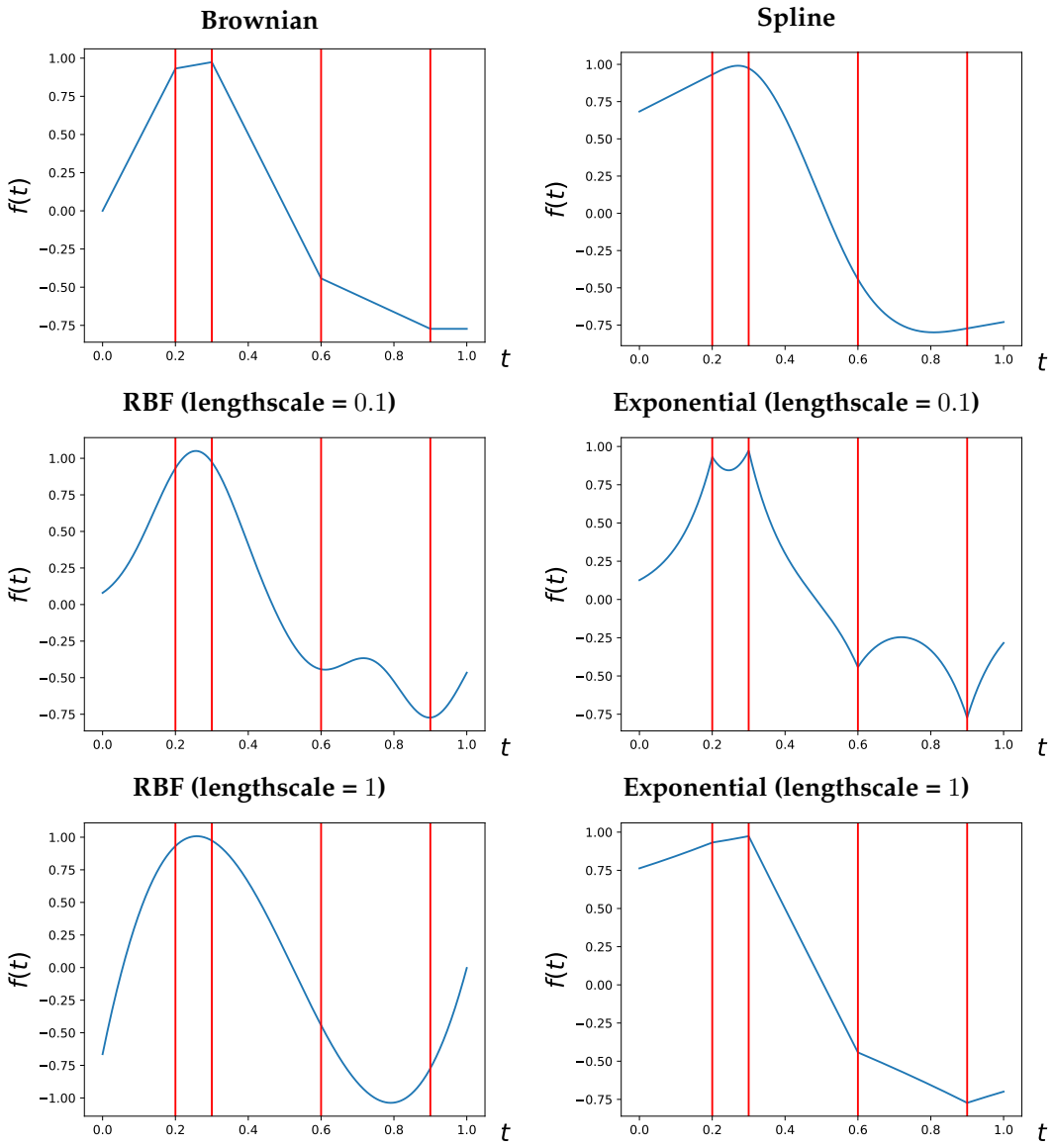

FIGURE 3.5: Interpolation between four values of a rescaled sine function in $[0, 1]$, given by Theorem 3.3.5 using different kernels. The points are marked with red vertical lines.

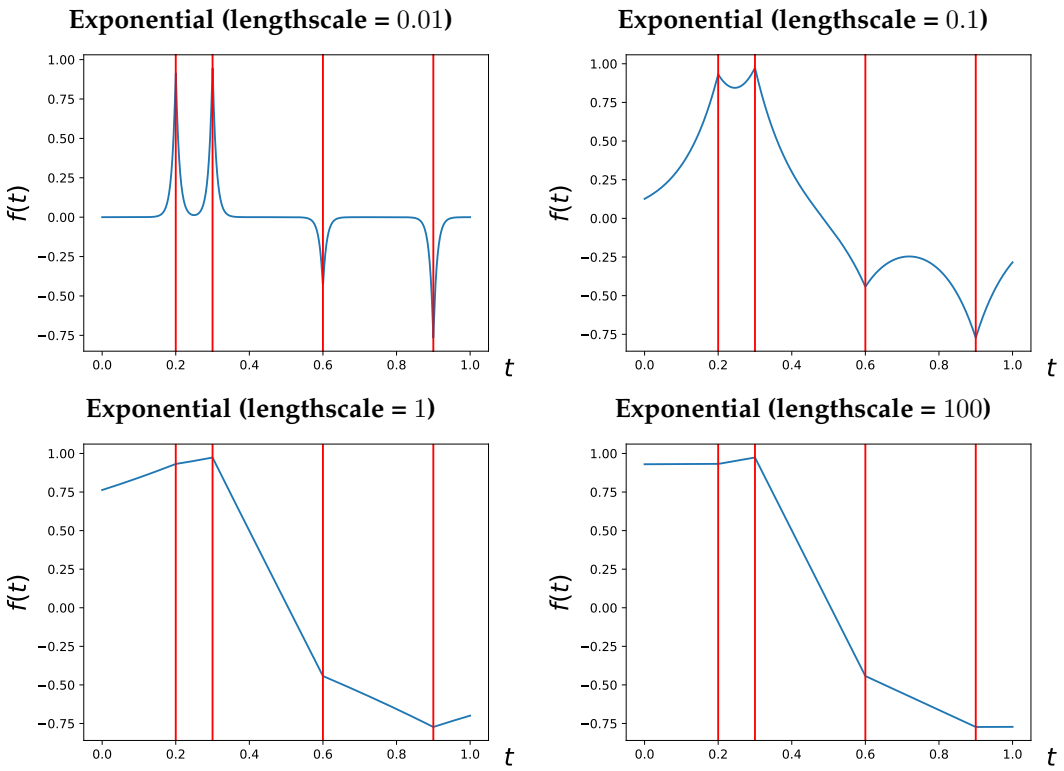

FIGURE 3.6: Interpolation between four values of a rescaled sine function in $[0, 1]$, given by Theorem 3.3.5 using the Ornstein-Uhlenbeck (exponential or Laplace) kernel with different lengthscales. The The lengthscale parameter of the exponential kernel determines the rate of reversion to the mean.

(iii) If $t < t_1$ then $f(t)$ is the linear interpolation between 0 and μ_1

$$f(t) = \mu_1 \left(\frac{t}{t_1} \right).$$

(iv) If $t > t_n$ then $f(t) = \mu_n$.

The proof of the above theorem is in Appendix B.

To summarize, in this chapter we have presented Recursive Maxima Hunting (RMH), a feature selection method for functional data. We have analyzed its mathematical properties, and we have found a relationship between RMH and interpolations of the class 1 mean.

In the next chapter, we will see that RMH using the Uniform-Brownian correction selects the features that appear in the Bayes rule when the noise process is the limit of an Ornstein-Uhlenbeck process whose lengthscale parameter tends to infinity.

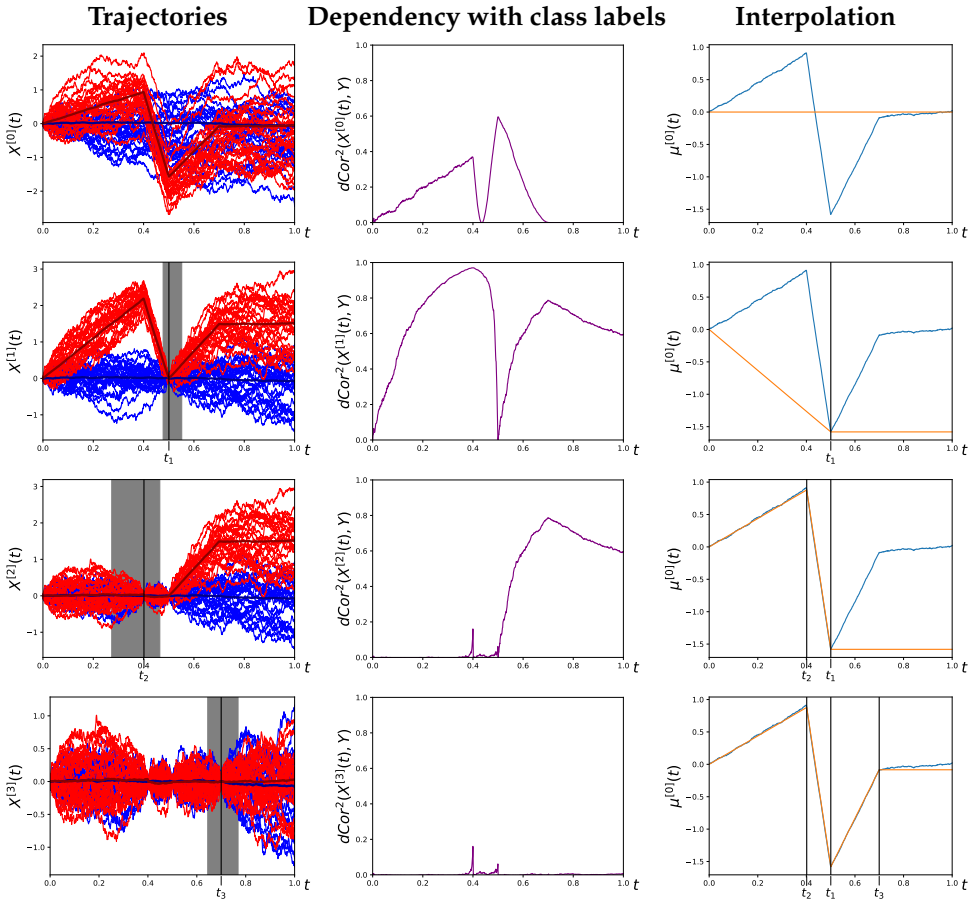

FIGURE 3.7: The plots presented in Figure 3.1 (first and second columns) are complemented by a third column of plots that illustrate how the correction process iteratively builds an approximation of the class 1 mean by interpolating between the values of this mean at the points selected by RMH. This third column shows, in blue, the original mean $\mu^{[0]}(t) = \mu(t)$ and, in orange, the interpolation given by $\mathbb{E}\left[Z^{[0]} \mid Z^{[0]}(t_1) = \mu^{[0]}(t_1), \ldots, Z^{[0]}(t_n) = \mu^{[0]}(t_n)\right]$ using the selected points and the Brownian kernel.

Chapter 4

RMH with the Uniform-Brownian correction

As described in the previous chapter, Recursive Maxima Hunting (RMH) is a variable selection method for functional data. It selects in each iteration the variable which maximizes a measure dependency with the class, and then corrects the observed trajectories subtracting its influence. As a consequence, RMH select variables that are not relevant by themselves, but are relevant together with the variables selected in previous iterations. In this chapter we will prove that RMH with the Uniform-Brownian correction will select the variables that appear in the Bayes rule when certain conditions are met.

We will assume throughout this chapter that in Equation (3.1), the class 1 mean $\mu(t)$ is a continuous piecewise linear function, which can be written as

$$\mu(t) = \begin{cases} \mu(s_0), & \text{if } t \in [0, s_0), \\ \mu(s_i)\left(1 - \frac{t-s_i}{s_{i+1}-t_i}\right) + \mu(s_{i+1})\left(\frac{t-s_i}{s_{i+1}-s_i}\right), & \text{if } t \in [s_i, s_{i+1}], \\ \mu(s_n), & \text{if } t \in (s_n, 1], \end{cases}$$

where $s_0 < \ldots < s_{i-1} < s_i < \ldots < s_n$ and the values $\mu(s_i)$ are arbitrary. We will also assume that the noise process $Z(t)$ verify the assumptions that the Uniform-Brownian correction makes; that is, $\mathbb{E}[Z \mid Z(t_{max}) = z] = z$, and upon conditioning on the value $Z(t_1)$, $Z \mid Z(t_1)$ becomes an isotropic standard Brownian process emanating from t_1. The covariance function of this process is

$$K(s,t) = \begin{cases} \min(|s-t_1|, |t-t_1|), & \text{if } (s-t_1)(t-t_1) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

A sample of trajectories of this process is shown in Figure 4.1. We will also assume that the noise process is stationary and Markovian. Although there is not any Gaussian process verifying these assumptions, in section 4.3 we will prove that a particular limit of an Ornstein-Uhlenbeck process verifies them. To determine relevance and redundancy of the features the squared distance covariance ($\mathcal{V}^2$) will be used as the dependency measure.

First we will illustrate the workings of a particular example for one kind of $\mu(t)$, and then we will present the general result for every $\mu(t)$ that is piecewise linear.

We will discuss later that there is a Gaussian process which, in a limit over its parameters, verify the conditions of the noise process declared above.

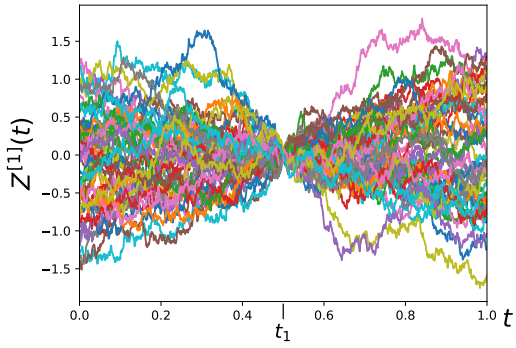

FIGURE 4.1: An isotropic Brownian process that emanates from $t_1 = 0.5$.

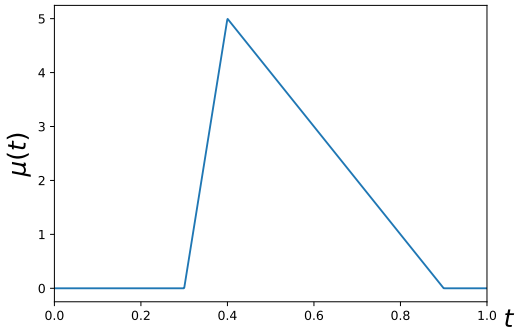

FIGURE 4.2: The mean used in the example. In this case, $s_1 = 0.3$, $s_2 = 0.4$, $s_3 = 0.9$ and $a = \mu(s_2) = 5$.

4.1 The Uniform-Brownian correction: a simple example

First, we will provide the derivation for a simple example that serves to illustrate how RMH selects the points in which the mean first derivative $\mu'(t)$ is discontinuous. Our problem will be one of the family explained in section 1.1. Suppose that in this case $\mu(t)$ is the piecewise linear function studied as an example in section 2.1.2

$$\mu(t) = \begin{cases} 0 \text{ if } t < s_1 & \\ a\frac{t-s_1}{s_2-s_1} & \text{if } s_1 < t < s_2 \\ -a\frac{t-s_2}{s_3-s_2} + a & \text{if } s_2 < t < s_3 \\ 0 \text{ if } t > s_3 & \end{cases}$$

where $a = \mu(s_2) > 0$ and $s_1 < s_2 < s_3$. As when the noise is Brownian, in this case the points that appear in the Bayes rule are s_1, s_2, and s_3. We will prove this affirmation in section 4.3, once we describe the process corresponding to the Uniform-Brownian correction. An example of this kind of function is shown in Figure 4.2.

We will now prove that in this case the first point selected by RMH is s_2. Since RMH selects the point that maximizes the dependency measure (in this case $\mathcal{V}^2$) with

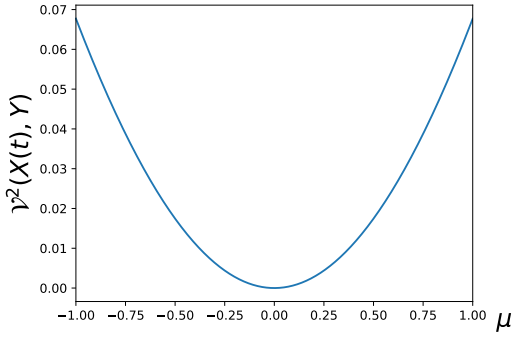

FIGURE 4.3: Here it has been represented the graph of $\mathcal{V}^2(X(t), Y)$ supposing $\sigma = 1$ as a function of μ. It is clear from this graph that $\mathcal{V}^2(X(t), Y)$ is a monotonically increasing function of $|\mu|$.

the class, we have to prove that s_2 is indeed the point where $\mathcal{V}^2(X(t), Y)$ has its maximum value. It is important to notice that, according to Corollary 2.2.2, $\mathcal{V}^2(X(t), Y)$ depends on both $\mu(t)$ and $\sigma(t) = \sqrt{K(t,t)}$. A consequence of this is that $\mathcal{V}^2$ is differentiable in all points where $\mu(t)$ and $\sigma(t)$ are differentiable. However, in our setting, $\mu(t)$ is non differentiable precisely at our points of interest s_1, s_2, s_3. Therefore, the maxima cannot be found using derivatives. The idea is to locate the maxima by determining whether the intervals between the points are increasing or decreasing.

The following theorem is useful to find the first point selected by RMH, as the noise process is stationary before any correction is made. The proof of the theorem can be found in Appendix B.

Theorem 4.1.1 (Maxima with constant variance)**.** Let X be a stochastic process that depends on a dichotomic variable Y. Let $X(t)$ be the random variable corresponding to X at time t and let $\mathcal{V}^2$ be the squared distance covariance function.

If X is a Gaussian process and $[X(t) \mid Y = 0]$, $[X(t) \mid Y = 1]$ are homoscedastic with constant standard deviation $\sigma(t) = \sigma$ and means 0 and $\mu(t)$ so that for every value of t

$$X(t) \mid Y = 0 \sim N(0, \sigma^2)$$
$$X(t) \mid Y = 1 \sim N(\mu(t), \sigma^2)$$

then $\mathcal{V}^2(X(t), Y)$ has the same increasing and decreasing intervals (and so, the same maxima) as $|\mu|$.

Since for a stationary Gaussian process $\sigma(t)$ is constant, the above theorem gives us the maxima of $\mathcal{V}^2(X(t), Y)$ if the noise process is stationary, as happens in our case. Also, $\mu(t)$ has only one maximum in s_2, so using Theorem 4.1.1, this will be the first point selected.

The first correction applied in RMH is then subtracting $\mathbb{E}[Z \mid Z(s_2) = \mu(s_2)] = \mu(s_2)$. We assume that the corrected noise process is a Brownian process emanating in both directions from the selected point, as shown in Figure 4.4. Therefore, the noise process is no longer stationary and Theorem 4.1.1 can not be applied again. We will now prove that RMH selects either s_1, on the interval to the left of s_2, or s_3, on the interval to the right of s_2. However, as discussed in subsection 3.3.3, since

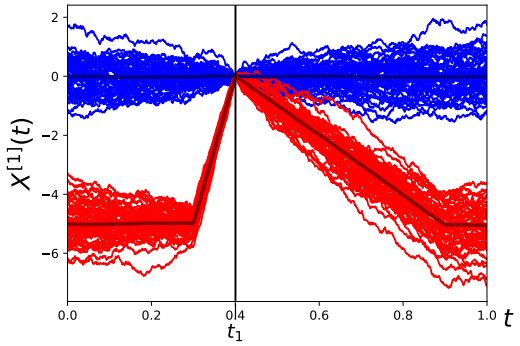

FIGURE 4.4: We assume that the corrected noise process after one point is selected is a Brownian process spawning in both directions from the selected point.

the process is Markovian it follows that the correction applied when we select either s_1 or s_3 will not affect the other interval. Thus, we can center our attention in the interval on the right-hand side of s_2 and show that RMH selects s_3 in this interval. A similar reasoning can be used to prove that RMH selects s_1 in the other interval.

First, we will prove that for $t > s_2$, $\mathcal{V}^2(X^{[1]}(t), Y)$ decreases monotonically with t. This can be done using the following theorem, which is proved in Appendix B.

Theorem 4.1.2 (Maxima with constant mean)**.** Let X be a stochastic process that depends on a dichotomic variable Y. Let $X(t)$ be the random variable corresponding to X at time t and let $\mathcal{V}^2$ be the squared distance covariance function.

If X is a Gaussian process and $[X(t) \mid Y = 0]$, $[X(t) \mid Y = 1]$ are homoscedastic with standard deviation $\sigma(t)$ and constant means 0 and μ so that for every value of t

$$X(t) \mid Y = 0 \sim N(0, \sigma(t)^2)$$
$$X(t) \mid Y = 1 \sim N(\mu, \sigma(t)^2)$$

then if μ is a nonzero constant $\mathcal{V}^2(X(t), Y)$ increases where $\sigma(t)$ decreases and vice versa. Therefore, the maxima of $\mathcal{V}^2(X(t), Y)$ are the minima of σ. If μ is zero then $\mathcal{V}^2(X(t), Y)$ is also zero.

Theorem 4.1.2 proves that the interval at the right side of s_3 is decreasing. We shall now prove that the interval between s_2 and s_3 is an increasing interval. We could try to use Corollary 2.2.2 and compute the first derivative as

$$\frac{d}{dt}\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[\frac{2\sigma'(t)}{\sqrt{\pi}}\left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1\right) + \mu'(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right) - 1\right)\right],$$

and prove that is positive on that interval. However, although is easy to see that the second term is positive, the first term is negative, and it is not clear that its absolute value is smaller than the second one. What we are going to do instead, is to use the information of the second derivative to show that the first derivative is positive.

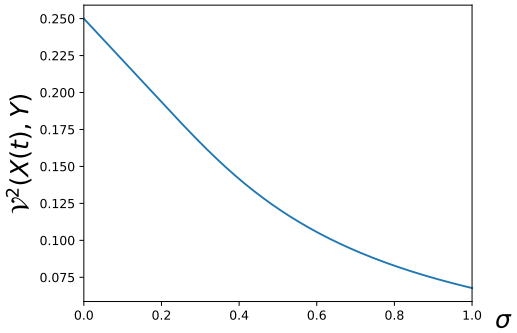

FIGURE 4.5: Here it has been represented the graph of $\mathcal{V}^2(X(t), Y)$ supposing $\mu = 1$ as a function of σ. It is clear from this graph that $\mathcal{V}^2(X(t), Y)$ is a monotonically decreasing function of σ.

We can now translate the origin to the point s_2, to simplify the formulas. In that case, we note that, in the interval between s_2 and s_3:

$$\begin{aligned}
\mu(t) &= kt \text{ with } k \text{ constant},\\
\mu'(t) &= k,\\
\sigma(t) &= \sqrt{t},\\
\sigma'(t) &= \frac{1}{2\sqrt{t}},\\
\frac{\mu(t)}{\sigma(t)} &= k\sqrt{t},\\
\lim_{t\to 0^+} 2\sigma'(t)\left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1\right) &= \lim_{t\to 0^+} \frac{\left(e^{-\frac{k^2 t}{4}} - 1\right)}{\sqrt{t}}\\
&= \lim_{t\to 0^+} \frac{-\frac{k^2}{4}e^{-\frac{k^2 t}{4}}}{\frac{1}{2\sqrt{t}}}\\
&= 0,\\
\lim_{t\to 0^+} \frac{\mu(t)}{\sigma(t)} &= 0.
\end{aligned}$$

Using the above results we can compute the value of the right-hand side of the first derivative $\frac{d}{dt}\mathcal{V}^2(X(t), Y)$ at the beginning of the interval.

$$\begin{aligned}
\lim_{t\to 0^+} \frac{d}{dt}\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \Bigg[&\frac{2\sigma'(t)}{\sqrt{\pi}}\left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1\right)\\
&+ \mu'(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right) - 1\right)\Bigg]\\
= 0.&
\end{aligned}$$

As the first derivative is non negative at the beginning of the interval, if the second derivative were positive, then the first derivative would increase and so, it would be positive on the remaining points. We will now show that this is indeed

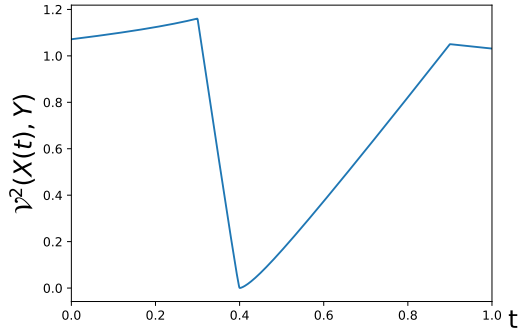

FIGURE 4.6: Squared distance covariance between points in the trajectory and the class once s_2 has been selected and the corresponding correction has been applied.

the case. Using again Corollary 2.2.2, we have the formula for the second derivative:

$$\frac{d^2}{dt^2}\mathcal{V}^2(X(t),Y) = 4p^2(1-p)^2\left[\frac{2\sigma''(t)}{\sqrt{\pi}}\left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}}-1\right)\right.$$
$$+\frac{1}{\sqrt{\pi}}e^{-\frac{\mu(t)^2}{4\sigma^2(t)}}\left(\frac{(\mu(t)\sigma'(t)-\sigma(t)\mu'(t))^2}{\sigma^3(t)}\right)$$
$$\left.+\mu''(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right)-1\right)\right].$$

Now, since $\mu''(t) = 0$ and $\sigma''(t) < 0$, the first two terms are positive and the remaining one is zero. Thus, the second derivative is positive. As said before, that means that the first derivative is positive and so the interval is an increasing one. As s_3 has a left increasing interval and a right decreasing one and the function is continuous at s_3, because is a composition of continuous functions, it follows that s_3 is the maximum of $\mathcal{V}^2(X(t), Y)$ in the interval at the right of s_2. The same reasoning can be applied to prove that s_1 is the maximum of $\mathcal{V}^2(X(t), Y)$ in the interval at the left of s_2. The graph of $\mathcal{V}^2(X(t), Y)$ can be seen in Figure 4.6.

After s_3 is selected, the RMH correction given by the interpolation function in Theorem 3.4.1 is applied. This correction consists in subtracting the function in $[s_2, 1]$ that is continuous, interpolates linearly between s_2 (the source of the Brownian process) and s_3 (the selected point), and is constant in $[s_3, 1]$. As can be seen in Figure 4.4, $\mu(t)$ has precisely this form after the correction corresponding to s_2 is applied. This means that the mean of the corrected trajectories will be 0 in the interval $[s_3, 1]$ for both clases. In consequence , according to expression Theorem 4.1.2, $\mathcal{V}^2(X(t), Y)$ will be 0 for every t in that interval.

A parallel argument can be made for the interval $[0, s_2]$: After having selected s_1 and applied the corresponding correction, the mean of the trajectories, and therefore $\mathcal{V}^2(X(t), Y)$, will be 0 also in the interval $[0, s_2]$. In consequence, the algorithm will stop and return the selected points s_1, s_2, s_3.

4.2 Uniform-Brownian: piecewise linear means

The analysis of the simple example presented in the previous sections suggests that, if the class 1 mean is piecewise linear, the RMH method with the Uniform-Brownian

correction selects the points at which the derivative of this mean is discontinuous, and, possibly, the extremes 0 and 1. We will now prove this conjecture in a general case.

The first step in the proof uses the same argument as in the simple example of the previous section. The stochastic process is initially assumed to be uniform and has a constant variance. Therefore, using Theorem 4.1.1, the first selected point will be the maximum of $\mu(t)$. Since $\mu(t)$ is piecewise linear, the maximum is either at the extremes of the interval $(0, 1)$, at a point at which the derivative $\mu'(t)$ is discontinuous, or belongs to a horizontal segment of degenerate maxima. In the last case, the point selected is not in the Bayes rule, but the rest of the argument stays the same, and the remaining selected points will be in the Bayes rule.

Since the process is Markovian, the selected point, t_1, splits the interval $[0, 1]$ into two subintervals that can be processed independently.

The first correction is then $\mu(t_1)$ and the mean of the class 1 corrected process is $\mu^{[1]}(t) = \mu(t) - \mu(t_1)]$. It is apparent that, after this first correction has been applied, $\mu^{[1]}(t)$ in each of the subintervals verifies Theorem 3.4.1: it is a piecewise linear function whose value at the selected point is 0. The corrected noise process is assumed to be an isotropic Brownian process that emanates from the selected point, as in Figure 4.1.

If we focus on the interval to the right of the selected point, after appropriate rescalings, the problem is equivalent to a binary functional classification problem with trajectories in $[0, 1]$ (i.e. of the same form as assumed in Equation (3.1)) in which the class 1 mean is piecewise linear and starts at zero. The noise is a standard Brownian process.

We will now show that, with these assumptions, RMH selects the points where the derivative of μ is discontinuous and, possibly, $t = 1$. By symmetry, similar arguments can be applied to the subinterval on the left-hand side of the selected point. The implication for the original process is that one selects the global maximum of $\mu(t)$, the points at which μ' is discontinuous, and possibly, the extremes, $t = 0$ and $t = 1$, and the first point, in case that the maximum of $\mu(t)$ belongs to an horizontal line.

The proof is based on showing that, under conditions that are somewhat more general than the ones assumed (piecewise linear mean, Uniform-Brownian correction), points at which μ' is differentiable cannot be maxima of $\mathcal{V}^2(X(t), Y)$. Therefore, only points at which the derivatives of $\mu(t)$ are discontinuous need to be considered as candidates for selection by RMH. This is stated in the following theorem, which is proved in Appendix B.

Theorem 4.2.1. With the same conditions that Corollary 2.2.2, if $\mathcal{V}^2(X(t), Y)$ is twice differentiable at point t, and

$$\mu''(t) = 0$$
$$\sigma''(t) < 0$$

then $\mathcal{V}^2(X(t), Y)$ does not have a maximum at point t, unless $\mathcal{V}^2(X(t), Y) = 0$ for every t.

Therefore, under these conditions, the maxima of the function $\mathcal{V}^2(X(t), Y)$ are at the points where $\mathcal{V}^2(X(t), Y)$ is not twice differentiable.

If $\mu(t)$ is a piecewise linear function, then $\mu''(t) = 0$ in the points at which $\mu(t)$ is differentiable twice. If the noise process is Brownian, $\sigma''(t) < 0$. Thus, using

Theorem 4.2.1 one concludes that points at which $\mu(t)$ is differentiable twice cannot be maxima of $\mathcal{V}^2(X(t), Y)$. The only possible maxima are either the extreme $t = 1$, or points at which $\mu'(t)$ is discontinuous.

The accumulated correction for a Brownian process after one or more points have been selected in RMH is given in Theorem 3.4.1. This has two consequences. First, the mean of the corrected process continues to be piecewise linear. Second, the non-differentiable points that have not been selected are still non-differentiable points. If at a point s, which is not at the extreme of the interval considered, the original class 1 mean $\mu(s)$ is not differentiable, then its left and right derivatives would be m_1 and m_2, with $m_1 \neq m_2$, corresponding to the slope of the straight lines on the left and the right of s, respectively. Now, if s has not been selected, the correction in a neighborhood of s is given by an interpolation between 0 and a selected point, the interpolation between two selected points or a constant value. In all of these cases, the correction consists in subtracting a straight line of slope m_3 in the neighborhood of s. Since differentiation is a linear operation, the new left and right derivatives at s will be $m_1 - m_3$ and $m_2 - m_3$, respectively, with $m_1 - m_3 \neq m_2 - m_3$. Thus, s continues to be non differentiable.

After one or more corrections have been applied, the Uniform-Brownian noise process becomes a set of Brownian bridges between 0 and the smallest selected point, and between consecutive selected points, and a Brownian process from the largest selected point and 1. For all these processes, $\sigma''(t) < 0$, which means that the demonstration presented applies at all the stages of the RMH algorithm.

Since $\mu(t)$ is piecewise linear, as in Theorem 3.4.1, the RMH algorithm halts when all the points at which $\mu'(t)$ is discontinuous (and possibly $t = 1$ are selected). At this point, the accumulated correction of the class 1 mean has the same shape that $\mu(t)$, as in the simple example described in the previous section.

4.3 The Uniform-Brownian process as a limit of the Ornstein-Uhlenbeck process

In the previous section we have shown that, when the mean $\mu(t)$ is piecewise linear, the Uniform-Brownian (UB) correction selects the points at which $\mu'(t)$ is discontinuous. In this section we provide some arguments to support use of the Uniform-Brownian correction.

In the original version of RMH, introduced in Torrecilla and Suárez, 2016, the noise process was assumed to be Brownian motion. A Brownian process is not stationary and assumes that all trajectories take the same value at 0. However, this is hardly the case in the real-world datasets. Therefore, one should probably use a noise process that does not single out any particular point in the trajectory. This can be accomplished assuming stationarity of the process.

The Gaussian process assumption is made so that explicit derivations can be made. Nevertheless, RMH with more general noise processes will be analyzed in future work.

A third property that is desirable besides stationarity and Gaussianity is the Markov property. This property is desirable because it allows to segment the selection task into independent subtasks.

The only Gaussian Markov process with continuous covariance function that is also stationary is the Ornstein-Uhlenbeck (OU) process, which corresponds to the exponential kernel (Bishop, 2006), as stated in the following theorem. The proof of this theorem is in Appendix B.

Theorem 4.3.1. Let Z be a Gaussian Markov and stationary process with a continuous covariance function K. Then,

$$K(s,u) = \sigma^2 \exp\left(-\frac{|s-u|}{l}\right),$$

where σ and l are constants. σ^2 is the variance of the process and $l > 0$ is the lengthscale parameter.

The Ornstein-Uhlenbeck is a mean-reverting process. In consequence, the approximation of the class 1 mean given by the accumulated corrections will tend to go to zero (the mean of the OU process) between the interpolation points selected by the RMH algorithm. As discussed in section 3.4, the lengthscale constant of the kernel controls how fast the process approaches towards the mean. A smaller lengthscale corresponds to a process with a faster reversion to the mean. In order that the noise process does not have a bias to a particular form for the mean, we assume that the lengthscale of the Ornstein-Uhlenbeck process tends to infinity. In this limit, the kernel K is constant. Therefore, the first correction of the RMH algorithm is

$$\mathbb{E}[Z \mid Z(t) = z] = z$$

as assumed in the UB process.

Using Lemma 3.3.6, it is possible to compute the RMH correction assuming that the noise is an OU process Assuming that the selected point is $t = 0$, the corrected noise is a zero-mean stochastic process whose covariance function is

$$\begin{aligned}
K_{new}(s,u) &= K(s,u) - \frac{K(s,t)K(t,u)}{K(t,t)} \\
&= \sigma^2 \exp\left(-\frac{|s-u|}{l}\right) - \frac{\sigma^2 \exp\left(-\frac{|s-t|}{l}\right)\sigma^2 \exp\left(-\frac{|t-u|}{l}\right)}{\sigma^2 \exp\left(-\frac{|t-t|}{l}\right)} \\
&= \sigma^2 \exp\left(-\frac{|s-u|}{l}\right) - \sigma^2 \exp\left(-\frac{|s-t|+|t-u|}{l}\right) \\
&= \sigma^2 \left[\exp\left(-\frac{|s-u|}{l}\right) - \exp\left(-\frac{|s|+|u|}{l}\right)\right]
\end{aligned}$$

Let us take the limit of this process as $l \to \infty$ and $\sigma^2 \to \infty$ with $\sigma^2 = \frac{1}{2}l$, so that the variance tends to infinity as the lengthscale tends to infinity:

$$K_{new}(s,u) = \lim_{l\to\infty} \frac{1}{2} l \left[\exp\left(-\frac{|s-u|}{l}\right) - \exp\left(-\frac{|s|+|u|}{l}\right)\right].$$

Replacing the exponential functions by their Taylor series expansions, we get

$$\begin{aligned}
K_{new}(s,u) &= \lim_{l\to\infty} \frac{1}{2} l \left[1 - \frac{|s-u|}{l} + O\left(\frac{1}{l^2}\right) - \left(1 - \frac{|s|+|u|}{l} + O\left(\frac{1}{l^2}\right)\right)\right] \\
&= \lim_{l\to\infty} \frac{1}{2} l \left[\frac{|s|+|u|-|s-u|}{l} + O\left(\frac{1}{l^2}\right)\right] \\
&= \lim_{l\to\infty} \frac{|s|+|u|-|s-u|}{2} + O\left(\frac{1}{l}\right) \\
&= \frac{|s|+|u|-|s-u|}{2}.
\end{aligned}$$

In this limit, if $s < t = 0 < u$ or $u < t = 0 < s$ then $K_{new}(s,u) = 0$. Also, if $t < s, u$ or $s, u < t$ then:

$$\begin{aligned} K_{new}(s,u) &= \frac{|s|+|u|-|s-u|}{2} \\ &= \begin{cases} \min(|s|,|u|) & \text{if } su > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

This is precisely the kernel of an isotropic Brownian process, as in Equation (4.1) that emanates from $t = 0$. Therefore, upon conditioning to a particular observed value, in the limit of large l, the OU process is a standard Brownian process that emanates from the point at which the process has been observed. The same reasoning can be applied assuming a different linear relation between lengthscale and variance, $\sigma^2 = Cl$, with C constant, to obtain a Brownian process with rescaled covariance.

Therefore, in this limiting sense, the Uniform-Brownian process becomes an isotropic Brownian process, once the correction at the first selected point has been applied. Since the conditional expectation for a Brownian process is a piecewise linear function starting at the origin From Theorem 3.4.1, we know that the interpolation given by the Uniform-Brownian process is a piecewise linear function between the selected points. The conditional expectation of the Brownian process becomes constant after the largest selected point. Thus, the Uniform-Brownian interpolation will also will also be constant to the right the largest selected point. By symmetry, it is also constant to the left the lowest selected point. In summary, the interpolation should be similar to the one depicted Figure 3.6 for a Ornstein-Uhlenbeck process with a large lengthscale.

Note that the conditional expectation is of the form analyzed in section 2.1.2

$$\begin{aligned} &\mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = \mu^{[0]}(t_1), \ldots, Z^{[0]}(t_n) = \mu^{[0]}(t_n)\right] \\ &= \begin{pmatrix} K(t,t_1) & \ldots & K(t,t_n) \end{pmatrix} \begin{pmatrix} K(t_1,t_1) & \ldots & K(t_1,t_n) \\ \vdots & \ddots & \vdots \\ K(t_n,t_1) & \ldots & K(t_n,t_n) \end{pmatrix}^{-1} \begin{pmatrix} \mu^{[0]}(t_1) \\ \vdots \\ \mu^{[0]}(t_n) \end{pmatrix} \\ &= \sum_{i=1}^{n} a_i K(t,t_i), \end{aligned}$$

for some real-valued constants $a_1, \ldots, a_n$. Therefore, for a functional classification problem in which the class 1 mean is piecewise linear and the noise is a Uniform-Brownian process, RMH selects the the points that appear in the Bayes rule, which, in this case, coincide with the points at which $\mu'(t)$ is not differentiable, and $t = 0$ and $t = 1$, if the function is not constant in the vicinity of these points.

We have shown in this chapter how RMH with the Uniform-Brownian correction can find the points that appear in the Bayes rule when the noise process correspond to a particular limit of Ornstein-Uhlenbeck processes whose lengthscale and variance tend to infinity and the mean is piecewise linear.

In the next chapter we will compare several variants of RMH with other dimensionality reduction methods, both in synthetic and real datasets.

Chapter 5

Empirical analysis of RMH

In this Chapter we present the results of exhaustive empirical of Recursive Maxima Hunting (RMH), the feature selection method for functional classification problems described and analyzed in chapter 3 and chapter 4. These experiments serve to evaluate the performance of RMH in synthetic and real-world problems. We will also present the results of experiments especially designed to illustrate some important properties of RMH. These experiments also serve to motivate future work on this method.

5.1 Empirical evaluation of RMH

In this section, we present the results of an exhaustive empirical evaluation. The study replicates and extend the experiments carried in Torrecilla and Suárez, 2016, including more dimensionality reduction methods and datasets in the comparison. Whenever possible the implementation used is taken from the widely used Python library Sklearn (Pedregosa et al., 2011). Otherwise they have been implemented anew in Python (MH, RMH, RK-VS and mRMR). Thus, these experiments serve as an independent replication of the results presented in Torrecilla and Suárez, 2016.

RMH is a filter feature selection process. Therefore, it is applied first to reduce the set of features to consider for the induction of the classifier from the training data. In all cases a k-nearest neighbors (k-NN) classifier has been used. The parameter k, the number of nearest neighbors, is chosen as an odd number in the range $[1, \sqrt{N_{train}}]$, where N_{train} is the number of trajectories in the training set, using 10-fold cross-validation. In most of the dimensionality reduction methods tested there is an additional parameter whose value has been selected among a grid of values by 10-fold cross-validation. In case of ties, the lower value for this parameter is selected.

The dimensionality reduction methods tested are:

- Not reducing the dimension at all, referred to as "Base".
- *Principal Component Analysis* (PCA), a multivariate analysis technique in which the vector of inputs are represented in the basis of eigenvectors of the sample covariance matrix (Hotelling, 1933). Once the change of basis is made (through an orthogonal transformation. That is, a rotation), one only keeps the components that correspond to the largest eigenvalues. The first principal component corresponds to the largest eigenvalue, the second one to the second largest and so on. Thus, the principal components selected in this order, yield the largest contribution to the variance of the original inputs. Dimensionality reduction can be achieved by keeping the initial components in the components in the sequence and discarding the rest. The components selected are linear combinations of the original ones. Therefore, they can be difficult to interpret.

We note that this method does not make use of the class labels in the classification problem, and so it does not necessarily select variables that have a high correlation with the class. Therefore, it is expected that this method performs worse in a classification problem that other methods that take into account the class labels, such as PLS.

In this work the Sklearn implementation of PCA is used and the number of components is chosen in the range $[1, 20]$ using cross-validation. The data is standardized as a previous step, as it is standard in this method.

- *Partial Least Squares* (PLS), a method that is similar to PCA, in the sense that it considers linear combinations of the original features, but takes into account the class labels. Specifically, it builds orthogonal features that are linear combinations of the original ones, and maximizes their covariance with the class labels (Rosipal and Krämer, 2006). This method does not have a unique implementation. There are several related algorithms that implement PLS variants using this basic idea (Wegelin et al., 2000). These algorithms could give different results for the same data. PLS has also been used in the context of FDA (Preda, Saporta, and Lévéder, 2007).

  We will use a widely used implementation of the PLS variant called PLS1, provided by Sklearn(PLSRegression). The number of components kept is chosen in the range $[1, 20]$ using cross-validation. This method also has its data standardized as a previous step.

- *Minimum Redundancy Maximum Relevance* (mRMR), a feature selection method in which one attempts to select a subset of the original variables that is jointly *relevant* and minimizes the *redundancy* among the selected variables (Ding and Peng, 2005). In this method the relevance of a subset $S = \{t_1, \ldots, t_n\}$ of features is defined as

$$\mathrm{Rel}(S) = \frac{1}{\mathrm{card}(S)} \sum_{t \in S} I(X(t), Y)$$

  where $I(\,\cdot\,,\,\cdot\,)$ is some *dependency measure* between random variables, like the ones described in section 2.2. The redundancy is estimated as

$$\mathrm{Red}(S) = \frac{1}{\mathrm{card}^2(S)} \sum_{s,t \in S} I(X(s), X(t)).$$

  Although mutual information is the original measure used to quantify the degree of dependence between these random variables, the method has also been tested with other measures of dependency, such as distance covariance or distance correlation (Berrendero, Cuevas, and Torrecilla, 2016a).

  The algorithm proceeds iteratively: The first feature selected in this method is the one that has the highest relevance. At each step, starting from a set of selected features S, it selects the unselected feature that maximizes either the difference $\mathrm{Rel}(S') - \mathrm{Red}(S')$ or the ratio $\mathrm{Rel}(S')/\mathrm{Red}(S')$, where S' is the set that includes S and the feature under consideration. Different rules can be used to halt the process of incorporating variables.

  We will use mRMR with Mutual Information as dependency measure, and the subtraction criterion, which has been reported to be more stable (Gulgezen,

Cataltepe, and Yu, 2009). In this method, the number of components is also chosen in the range $[1, 20]$ using cross-validation.

- *Reproducing Kernel Variable Selection* (RK-VS) method, a recent method of feature selection in the context of binary classification (Berrendero, Cuevas, and Torrecilla, 2017) in which the Mahalanobis distance between the multivariate means corresponding to the selected variables of the two classes is maximized. The method assumes that the classification problem is of the form assumed in in Equation (2.1)

$$X(t) = \begin{cases} Z(t) & \text{if } Y = 0 \\ \mu(t) + Z(t) & \text{if } Y = 1, \end{cases}$$

where $Z(t)$ is a zero-mean Gaussian process with kernel K. The mean $\mu(t)$ verifies the sparsity assumption section 2.1.2

$$\mu(\,\cdot\,) = \sum_{i=1}^{n} a_i K(\,\cdot\,, t_i).$$

The method also assumes that the number n of points that appear in the Bayes rule is known. Under these assumptions, the Bayes error is a monotone decreasing function of $||\mu||_K$, the norm of $\mu(t)$ in $\mathcal{H}(K)$, the Reproducing Kernel Hilbert space associated with kernel K. The square of $||\mu(t)||_K$ can be expressed as

$$\begin{aligned} ||\mu||_K^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(t_i, t_j) \\ &= \begin{pmatrix} \mu(t_1) & \dots & \mu(t_n) \end{pmatrix} \begin{pmatrix} K(t_1,t_1) & \dots & K(t_1,t_n) \\ \vdots & \ddots & \vdots \\ K(t_n,t_1) & \dots & K(t_n,t_n) \end{pmatrix}^{-1} \begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_n) \end{pmatrix}, \end{aligned}$$

which correspond to the Mahalanobis distance between the vectors of means evaluated at the specified points. The method thus tries to select the points $t_1, \dots, t_n$ that maximize this quantity once $\mu(\,\cdot\,)$ and $K(\,\cdot\,,\,\cdot\,)$ are replaced with their natural estimators. That is done using a greedy approach that selects one point at a time.

This method performed very well in previous experiments with functional data, and was compared with *Maxima Hunting* (Torrecilla, 2015).

For the experiments in this work, the number of components is also chosen in the range $[1, 20]$ using cross-validation.

- Maxima Hunting (MH). This method has been described in section 2.3.

  In this method *distance correlation* is used as dependency measure. To identify local maxima some smoothing is applied: A point t is considered a local maximum only if it is a global maximum in an interval that includes h discretization points both to the left and to the right of t. The parameter h is here chosen by the cross-validation grid search in the range $[1, 10]$.

- Recursive Maxima Hunting (RMH(S)), the method described in this work (chapter 3), with a GP correction that uses as covariance function the sample covariance of the trajectories (after subtracting the means of each class).

The value of the parameter `min_relevance` is fixed at 0.8. The parameter `min_redundancy` is chosen among the values 0.025, 0.05 and 0.1 using 10-fold cross-validation. For a fair comparison with the other methods, at most 20 components are selected. The dependency measure used is *distance correlation*.

- Recursive Maxima Hunting (RMH(B)) with a standard Brownian correction, as in Torrecilla and Suárez, 2016 and the same parameters as RMH with the correction using the sample covariance.
- Recursive Maxima Hunting (RMH(U-B)) with a Uniform-Brownian correction, and the same parameters as RMH with the correction using the sample covariance.

5.1.1 Experiments on synthetic data

The different dimensionality reduction methods considered have been tested on synthetic data generated from the model

$$X(t) = \begin{cases} B(t) \\ B(t) + \mu(t) \end{cases}$$

where $B(t)$ is the standard Brownian process and $\mu(t)$ is one of the following means:

- peak: the function $\mu(t) = 2\Phi_{3,3}(t)$, where

$$\Phi_{m,k}(t) = \int_0^t \sqrt{2^{m-1}} \left[\mathbb{I}_{\left(\frac{2k-2}{2^m}, \frac{2k-1}{2^m}\right)} - \mathbb{I}_{\left(\frac{2k-1}{2^m}, \frac{2k}{2^m}\right)} \right]$$

  This mean corresponds to a piecewise linear function of the family studied in section 2.1.2. Thus, when this mean is used, the Bayes rule depends only on the points $X\left(\frac{1}{2}\right)$, $X\left(\frac{5}{8}\right)$, and $X\left(\frac{3}{4}\right)$. The Bayes error in this case is $L^* \simeq 0.1587$.

- peak2: the function $\mu(t) = 2\Phi_{3,2}(t) + 3\Phi_{3,3}(t) - 2\Phi_{2,2}(t)$, using the previous definition of $\Phi_{m,k}(t)$. This is a linear combination of functions of the family studied in section 2.1.2. The resulting function is no longer a member of this family. However, it is a piecewise linear function of the form Theorem 3.4.1. Thus, it verifies the sparsity assumption described in section 2.1.2 with the Brownian kernel. Therefore, the only points that appear in the Bayes rule are $X\left(\frac{1}{4}\right)$, $X\left(\frac{3}{8}\right)$, $X\left(\frac{1}{2}\right)$, $X\left(\frac{5}{8}\right)$, $X\left(\frac{3}{4}\right)$, and $X\left(1\right)$. The Bayes error in this case is $L^* \simeq 0.0196$.

- square: the function $\mu(t) = 2t^2$. In this case, the Bayes rule depends on the whole trajectory. Nonetheless, the Bayes error can be computed and is $L^* \simeq 0.1241$.

- sin: the function $\mu(t) = \frac{1}{2}\sin(2\pi t)$. Bayes rule does not depend in a finite number of points in this case either. The Bayes error is $L^* \simeq 0.1333$.

Simulations of trajectories for each of these problems are shown in Figure 5.1.

As in the experiments carried out in Torrecilla and Suárez, 2016, training sets with sizes 50, 100, 200, 500 and 1000 and test sets of size 1000 were generated. The trajectories were discretized in a grid of 200 regularly-spaced points.

The results of the experiments on synthetic data are summarized in Figure 5.2. The values reported are averages over 200 realizations of the training and test sets

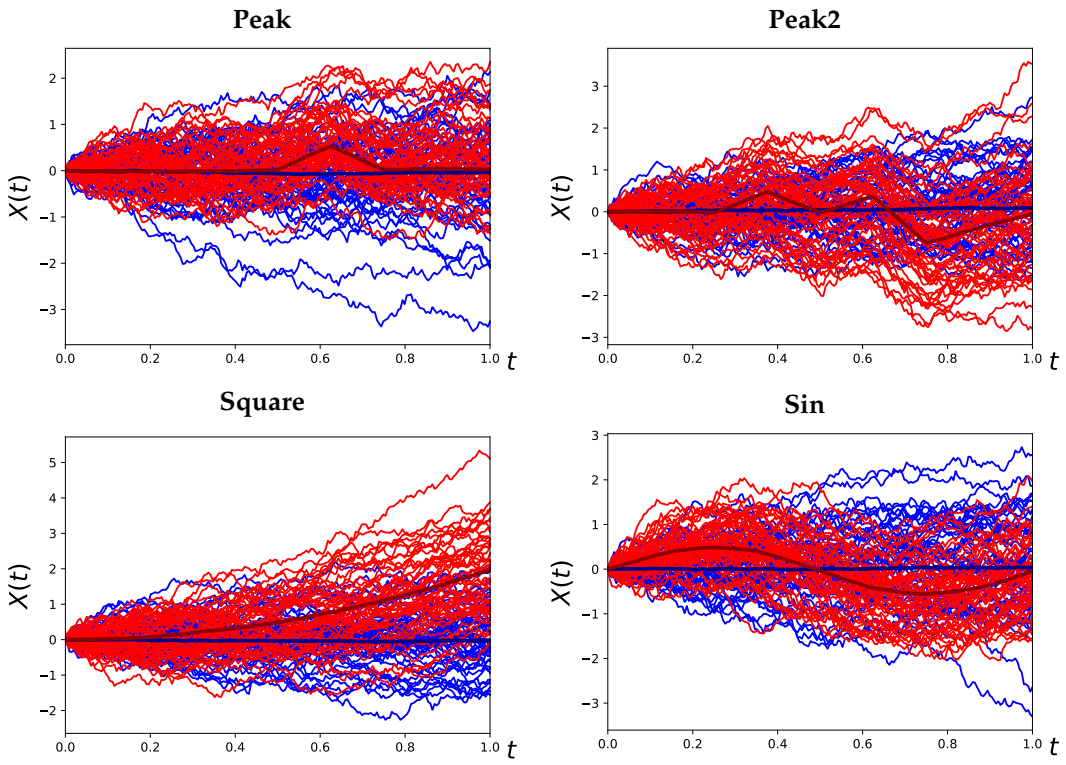

FIGURE 5.1: Simulations of trajectories for the synthetic classification problems considered.

for each of the problems. The plots in the first row display the error of each method as a function of the size of the training set for the different classification problems considered. The Bayes error is marked with a horizontal dashed line. The plots in the second row of this figure displays the number of variables selected by each method as a function of the size of the training set. "Base" is not included in these plots in the second row because this method makes use of all the variables.

From these results one can see that, as expected, PCA is the worst dimensionality reduction method, because it does not take into account the class labels. The overall performance of "Base" is rather poor. This means that the dimensionality reduction step is useful to build accurate classifiers.

Even though most results obtained are similar to those presented in Torrecilla and Suárez, 2016, there are some differences related to the choices made in the implementation of the different methods. For instance, the number of features selected by the MH method, specially with the "square" mean, is significantly different than that of the original study. This is probably related to the way in which ties are resolved in cross-validation grid search for the value of the smoothing parameter, which is different from the one implemented by Sklearn. The current implementation of PLS yields better results than in Torrecilla and Suárez, 2016.

The best overall performance corresponds to the RMH methods. They are both accurate and select small numbers of variables. Among the RMH methods, the worst performance corresponds to using the sample covariance. This poorer performance arises from sampling errors in the estimation of the covariance function for the noise process. In particular, the accuracy of this methods is poorer for small training sizes. Furthermore, too many variables are selected. This methods always select at least

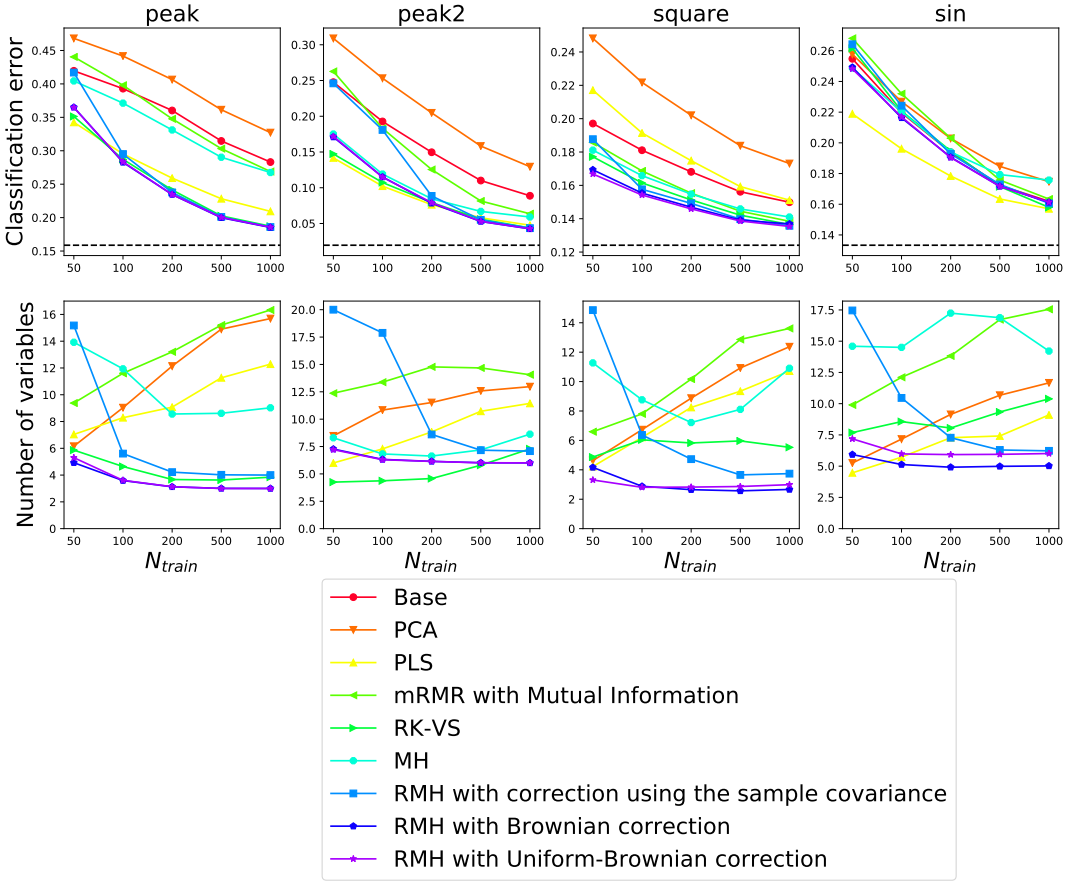

FIGURE 5.2: The results of the experiments on synthetic classification problems. The Bayes error is shown with a dashed line.

one more variable than RMH with a Brownian correction. The reason is that covariance estimated using the sample is not exact, and the correction does not left the point $t = 0$ fixed. Thus, that point will later be selected, as we can see in Appendix D.

The accuracies of RMH with the Brownian correction and the Uniform-Brownian correction are almost identical in all cases. This should be expected, since they become identical after the first iteration of the latter is made. A subtle detail is that, in the "sin" case, the Uniform-Brownian correction selects one more variable than the Brownian correction. That is because the sin function is zero at the origin. Therefore, RMH with the Uniform-Brownian correction selects that point, to have a better interpolation of the mean, as explained in section 3.4. By contrast, RMH with the Brownian correction does not select that point, because, by construction, every interpolation with the Brownian kernel begins at zero. therefore, the only disadvantage of using the Uniform-Brownian correction when the actual noise process is Brownian is, at most, selecting one more variable.

5.1.2 Experiments on real-world data

The dimensionality reduction methods have been evaluated also with real-world functional classification problems. The discretization points of those datasets have been reshaped to be points in the interval $[0, 1]$ with a uniform separation. The

datasets have been partitioned at random in a *train set*, with $\frac{2}{3}$ of the observations, and a *test set*, with $\frac{1}{3}$ of the observations, in a stratified way. This partitioning process has been done 200 times. The results reported are averages over these different partitions. The mean and standard deviation of the accuracy score and the number of variables selected for each method and dataset are shown in Table 5.1 and Table 5.2 respectively. Also, the plots of each dataset and box plots of the error and number of features selected for every method in each one can be found in Appendix C. The results present small deviations from Torrecilla and Suárez, 2016. These can be explained because the strategy used by Sklearn to resolve ties in the grid search cross validation is different from the one used in the original study. Thus, the optimal parameters selected could differ in case of ties.

The datasets considered in the experiment are:

- The *Berkeley Growth Study* (Figure C.1). In this problem the curves correspond to the heights of 54 girls an 38 boys (Ramsay, 2006; Mosler and Mozharovskyi, 2015). The observations are discretized at 31 non equidistant ages between 1 and 18. The RMH methods have very good performance. This example shows also that, when the trajectories do not begin at 0, the Brownian correction selects many unnecessary variables that are not selected by the Uniform-Brownian correction. Thus, in this example RMH with the Uniform-Brownian shows an improvement over the Brownian correction. Also both of them perform better than RMH with a correction based on the sample covariance.

- The *Tecator dataset* (Figure C.2). This dataset consists of 215 near-infrared absorbance spectra of finely chopped meat (Ferraty and Vieu, 2006; Galeano, Joseph, and Lillo, 2015). The trajectories are sampled at 100 equally spaced points. The class labels correspond to the level of fat content being above or below 20%. The trajectories used correspond to the second derivative of the original curves, as is recommended (Ferraty and Vieu, 2006). An important observation in this example is that, although the RMH methods have good performance, RMH with the Brownian or Uniform-Brownian correction select the maximum allowed number of variables. This happens because they attempt to select enough variables to give a good interpolation of the mean difference, even if they are unnecessary to have a good performance on the classification problem. This suggests that RMH could be improved changing the stopping criterion, so that it does not select more relevant variables if they do not significantly reduce the classification error.

- The *Phoneme dataset* (Figure C.3). The trajectories in this dataset are 1717 log-periodograms constructed from 32 millisecond long recordings of males pronouncing two phonemes: the phoneme "aa"(695 curves) and the phoneme "ao"(1022 curves). Those trajectories are discretized at 256 equidistant points (Galeano, Joseph, and Lillo, 2015). Following Ferraty and Vieu, 2006, the data has been smoothed and truncated to the first 150 features. In this example we can clearly see how the RMH methods have better performance than most of the other methods and generally select fewer variables.

- The *Medflies dataset* (Figure C.4). In this dataset the trajectories correspond to daily egg-laying patterns of flies. They are 512 30-day curves (beginning at day 5) of flies which live at most 34 days and 266 curves of long-lived flies (that reach the day 44) (Mosler and Mozharovskyi, 2015). This example also shows how the RMH methods achieve better results than the other methods with fewer variables.

Accuracy score									
	Base	PCA	PLS	mRMR(MI)	RK-VS	MH	RMH(S)	RMH(B)	RMH(U-B)
Berkeley	*0.956 ± 0.032*	0.946 ± 0.035	**0.958 ± 0.032**	0.942 ± 0.034	0.939 ± 0.038	0.938 ± 0.043	0.935 ± 0.044	0.950 ± 0.033	0.942 ± 0.034
Tecator	0.979 ± 0.015	0.987 ± 0.017	0.980 ± 0.017	0.982 ± 0.016	0.978 ± 0.017	*0.988 ± 0.014*	**0.990 ± 0.011**	0.985 ± 0.013	0.987 ± 0.012
Phoneme	0.796 ± 0.015	0.798 ± 0.015	*0.810 ± 0.014*	0.801 ± 0.015	**0.812 ± 0.013**	0.805 ± 0.013	0.801 ± 0.014	0.809 ± 0.013	0.809 ± 0.014
Medflies	0.544 ± 0.031	0.550 ± 0.032	0.583 ± 0.038	0.600 ± 0.038	0.595 ± 0.035	0.590 ± 0.036	0.599 ± 0.037	**0.604 ± 0.037**	*0.603 ± 0.036*
Gun	**0.936 ± 0.027**	0.924 ± 0.037	0.921 ± 0.035	*0.926 ± 0.036*	0.899 ± 0.041	0.909 ± 0.032	0.908 ± 0.038	0.925 ± 0.040	0.925 ± 0.041
MCO	0.819 ± 0.066	0.821 ± 0.069	*0.887 ± 0.058*	0.884 ± 0.061	0.884 ± 0.054	0.848 ± 0.066	**0.888 ± 0.065**	0.877 ± 0.059	0.875 ± 0.060
Coffee	0.984 ± 0.029	0.935 ± 0.061	0.984 ± 0.033	0.964 ± 0.045	0.968 ± 0.051	0.978 ± 0.035	0.956 ± 0.057	**0.995 ± 0.015**	*0.994 ± 0.016*

TABLE 5.1: Accuracy score for each method in the real dataset. The numbers shown are the mean and standard deviation. The greater mean is show in bold, and the second greater is italicized.

Number of selected variables									
	Base	PCA	PLS	mRMR(MI)	RK-VS	MH	RMH(S)	RMH(B)	RMH(U-B)
Berkeley	31.000 ± 0.000	*3.115 ± 1.733*	**2.560 ± 1.564**	5.275 ± 3.110	6.105 ± 4.608	3.960 ± 1.062	4.010 ± 1.162	6.570 ± 1.576	4.025 ± 0.254
Tecator	100.000 ± 0.000	**2.030 ± 0.699**	*2.060 ± 1.291*	5.095 ± 5.334	5.970 ± 5.880	5.375 ± 4.369	2.620 ± 1.255	20.000 ± 0.000	20.000 ± 0.000
Phoneme	150.000 ± 0.000	9.025 ± 4.399	8.980 ± 4.605	10.085 ± 5.581	6.900 ± 3.294	4.205 ± 0.658	**1.060 ± 0.341**	*3.890 ± 1.071*	5.050 ± 1.199
Medflies	30.000 ± 0.000	6.260 ± 4.349	6.885 ± 5.028	3.550 ± 2.565	3.920 ± 3.718	3.835 ± 2.381	**1.330 ± 0.501**	*1.620 ± 0.596*	1.645 ± 0.607
Gun	150.000 ± 0.000	15.810 ± 3.311	10.400 ± 3.796	10.895 ± 5.255	9.595 ± 4.578	**7.305 ± 2.912**	*8.315 ± 3.223*	10.565 ± 2.539	10.470 ± 2.404
MCO	360.000 ± 0.000	7.545 ± 2.590	*5.985 ± 2.628*	9.980 ± 4.646	**4.085 ± 2.300**	12.615 ± 8.332	7.605 ± 4.271	18.695 ± 1.934	18.535 ± 2.262
Coffee	286.000 ± 0.000	5.070 ± 2.560	**2.060 ± 1.103**	*3.945 ± 2.143*	4.075 ± 4.209	13.540 ± 5.484	4.540 ± 3.899	20.000 ± 0.000	20.000 ± 0.000

TABLE 5.2: Number of features selected for each method in the real dataset. The numbers shown are the mean and standard deviation. The smaller mean is show in bold, and the second smaller is italicized.

- The *Gun dataset* (Figure C.5). The dataset comes from the video surveillance domain (Ratanamahatana and Keogh, 2004). The datasets consists of 200 trajectories, each discretized to 150 points. Each one stores the position in the x axis of the centroid of the right hand of an actor while the actor is performing one of two possible actions. In 100 trajectories the actor draws a gun and points it at a target. In the other 100 trajectories the actor points at the target with the index finger. The classification problem is consists in detecting whether a person is drawing a real gun or simply imitating this gesture. In this case performance of the RMH methods is average, and there is no clear winner. In this example, the methods that achieve lower error do so by selecting more variables.

- The *MCO dataset* (Figure C.6). The trajectories are the measures of the mitochondrial calcium overload (MCO); that is, the level of the ion Ca^{2+} (Ruiz-Meana et al., 2003; Cuevas, Febrero, and Fraiman, 2004; Cuevas, Febrero, and Fraiman, 2006; Baíllo, Cuevas, and Cuesta-Albertos, 2011). This variable was observed every 10 seconds during an hour in isolated mouse cardiac cells. The aim of the study was to assess whether a drug called Cariporide increased the MCO level. The data has a control group with 45 trajectories and a group with 44 trajectories treated with Cariporide. This is another example where the RMH methods using Brownian and Uniform-Brownian corrections select too many variables compared with the others. Also, the function that measures the distance correlation between each feature and the class has many local maxima. Thus, MH selects many more variables than the other methods, as this method is the only one not restricted to select at most 20 features.

- The *Coffee dataset* (Figure C.7). This data has 56 spectrograms that belong to the Arabica and Robusta coffee variants (Briandet, Kemsley, and Wilson, 1996; Bagnall et al., 2012). Each trajectory is discretized to 286 points. The objective is to assign to each spectrogram its corresponding coffee variant. This example illustrates the near-perfect classification phenomenon described in subsection 2.1.3. In this example most methods can select enough features to achieve zero classification error in most cases. However, the RMH methods still select too many features, as they try to obtain a good interpolation of the mean difference.

5.2 Other experiments

The results of the following experiments serve to illustrate some interesting properties about RMH. These properties will be explored in more depth in future work.

5.2.1 Uniform-Brownian with different noise processes

In chapter 4 we determined the points that RMH with the Uniform-Brownian correction selects when the mean of the second class is a piecewise linear function. However, the only property of the noise process that we needed is that the second derivative of its standard deviation is less than zero ($\sigma''(t) < 0$). This suggests that the Uniform-Brownian process will select the same points even if the underlying noise process is different from the one assumed (the Uniform-Brownian process).

To test this hypothesis, we have tested a simple example with a piecewise linear mean and several noise processes, where the correction used by RMH is the

Uniform-Brownian one. In Figure 5.3, Figure 5.4, Figure 5.5 and Figure 5.6 we can see the examples where the noise processes have, respectively, a Matern $3/2$ kernel, a RBF kernel with lengthscale 0.1, a RBF kernel with lengthscale 1 and an exponential kernel with lengthscale 1. RMH with the Uniform-Brownian correction has selected for each example the same points that we proved that it should select when the noise process is the right one. However, the noise processes are very different, and the function that measures the dependency with the class has very different shapes for each of the examples (in the RBF example with lengthscale 0.1 it is even possible to miss a point if the minimum_relevance threshold is too low). Looking at these empirical results, we should say that the process used for the correction has greater importance that the real noise process for selecting the points in RMH.

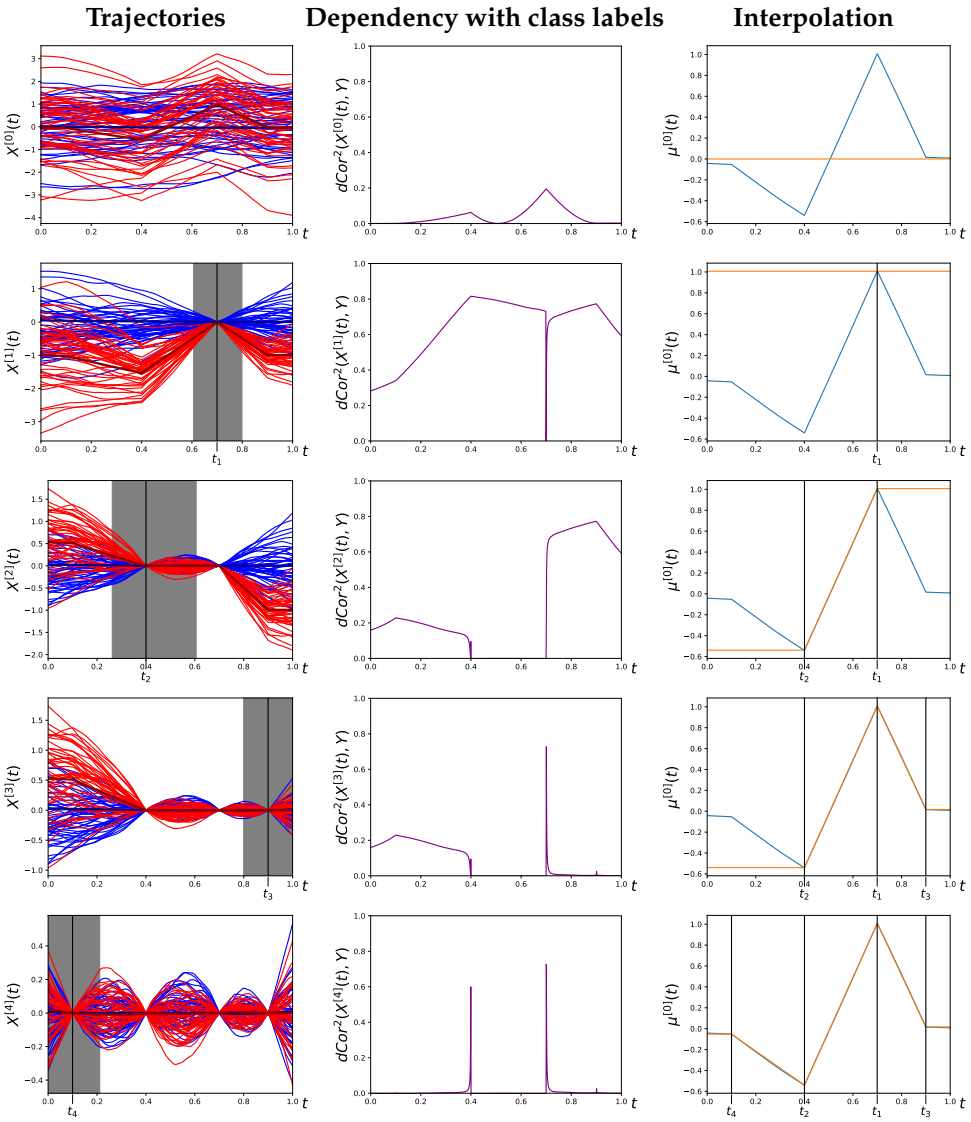

FIGURE 5.3: Example of RMH with the Uniform-Brownian correction applied to a problem where the noise process has a Matern 3/2 kernel.

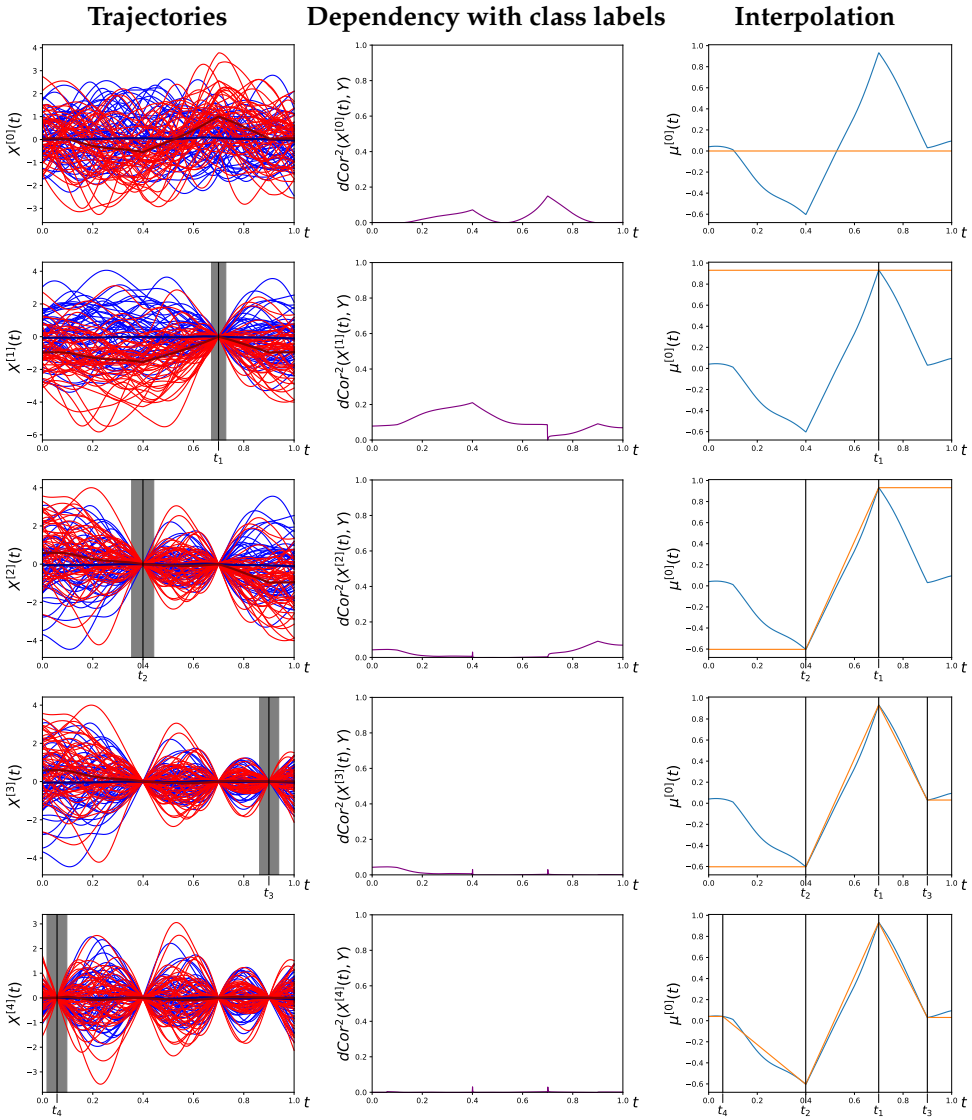

FIGURE 5.4: Example of RMH with the Uniform-Brownian correction applied to a problem where the noise process has an RBF kernel with lengthscale 0.1.

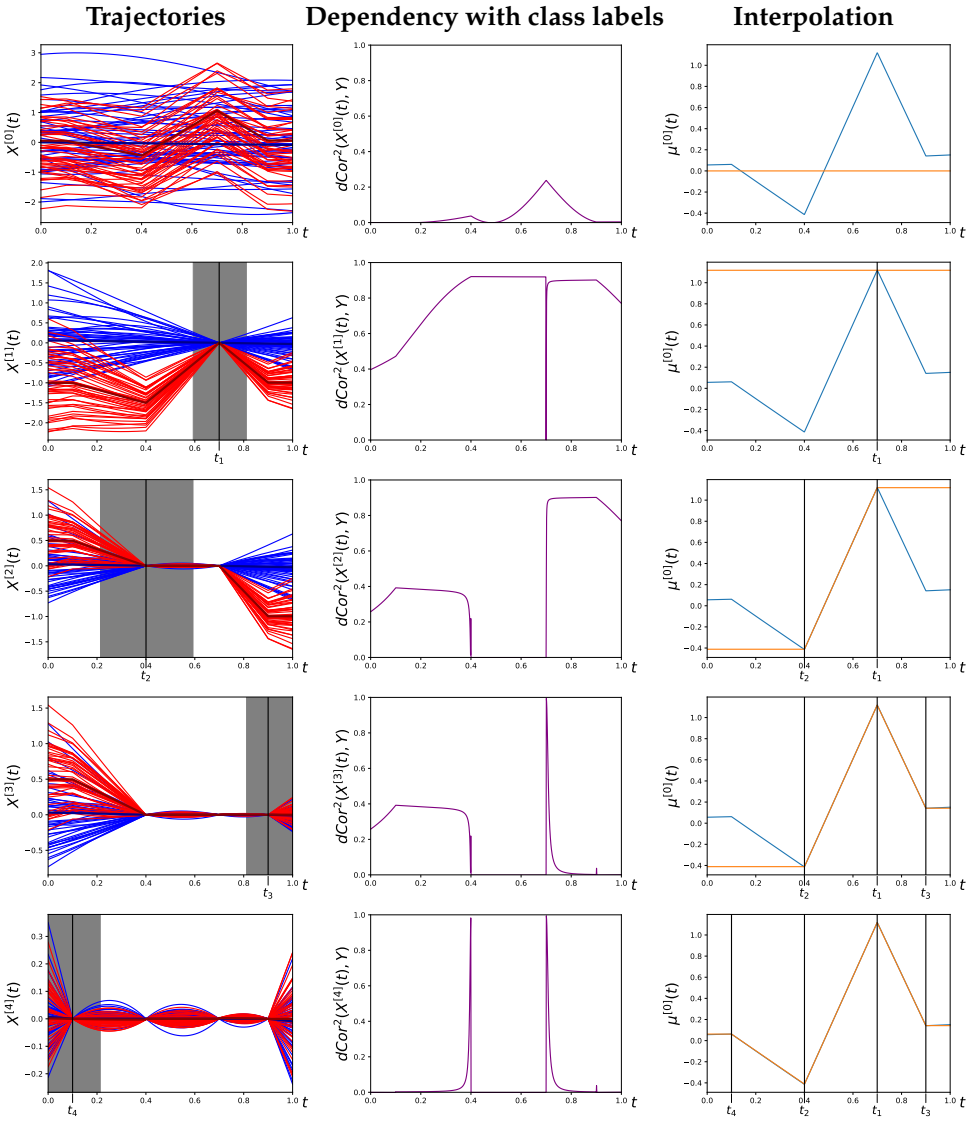

FIGURE 5.5: Example of RMH with the Uniform-Brownian correction applied to a problem where the noise process has an RBF kernel with lengthscale 1.

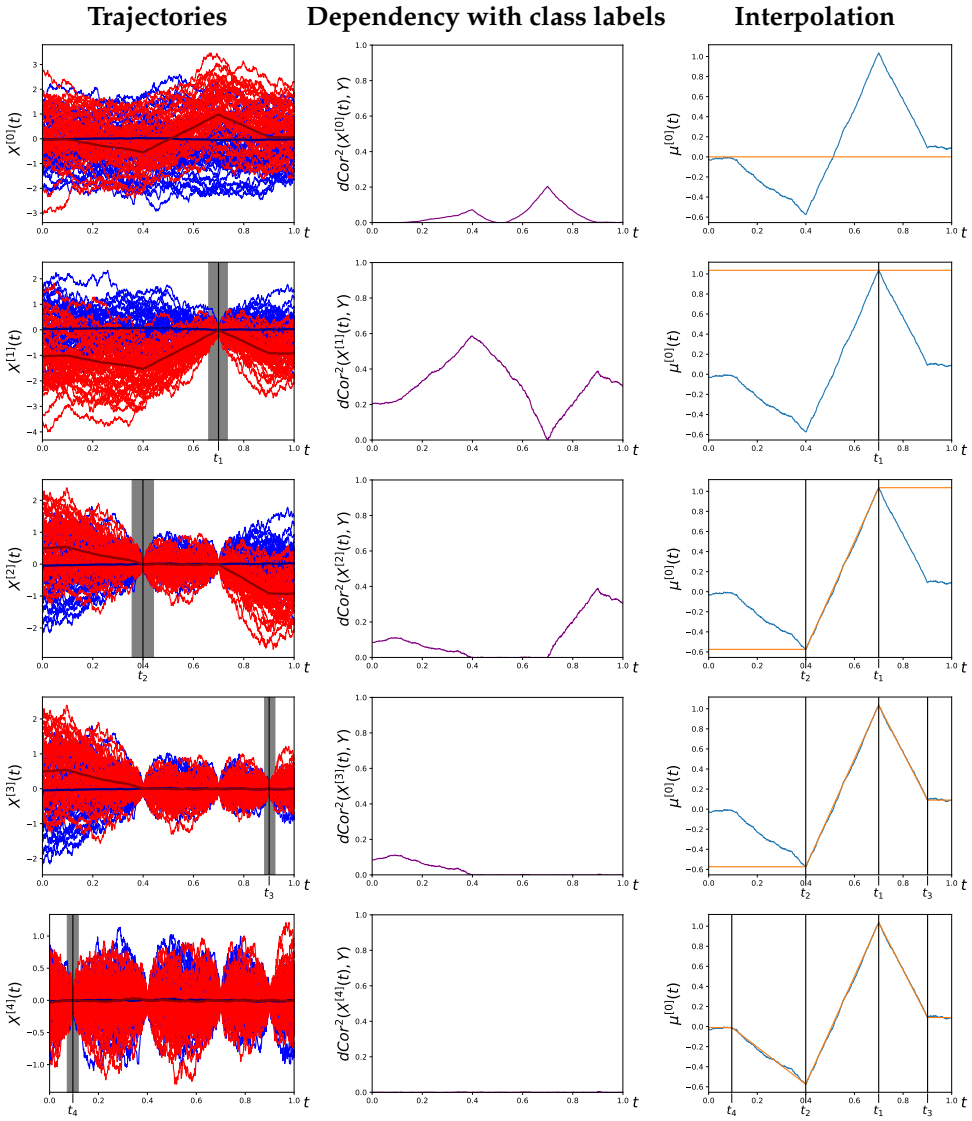

FIGURE 5.6: Example of RMH with the Uniform-Brownian correction applied to a problem where the noise process has an exponential kernel with lengthscale 1.

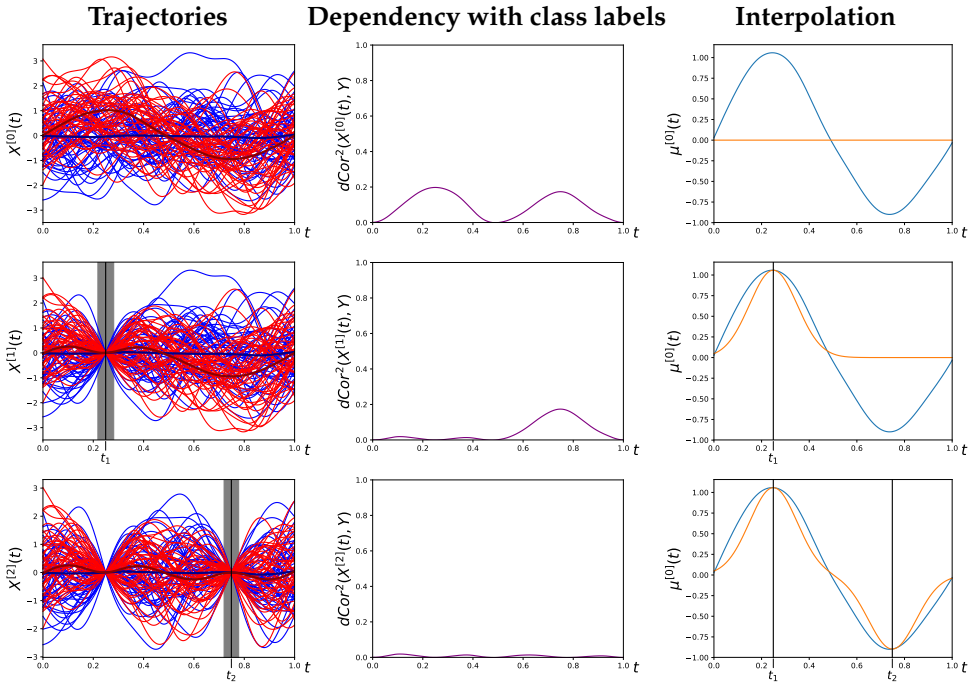

FIGURE 5.7: Example of RMH with a GP correction and a RBF kernel. The noise process has also a RBF kernel and the mean is in the associated RKHS, so the processes are equivalent.

5.2.2 Near-perfect classification

We now explore the behavior of RMH when presented with a problem in which the measures of the trajectories on each of the two classes are mutually singular. In this case, the problem has near-perfect classification (Delaigle and Hall, 2012), as explained in subsection 2.1.3. Specifically, we will first apply RMH with a GP correction with the RBF kernel. The RKHS associated to the RBF kernel consists in the functions that can be expressed as a power series that converge in $\mathbb{R}$ (Steinwart, Hush, and Scovel, 2006). In Figure 5.7 we show the iterations of RMH for a problem where the class 1 mean is a sine function which is in the RKHS associated to the RBF kernel. Therefore the two processes are equivalent. We can see that, as in all executions of RMH shown up to this point, the trajectories tend to look more similar after the corrections have been made. In Figure 5.8 we show another problem in which the class 1 mean is a piecewise linear function. This mean can not be expressed as a power series that converge in $\mathbb{R}$. Therefore, it does not belong to the RKHS associated to the RBF kernel. In consequence, this problem is an example where the measures of the classes are mutually singular and we have near-perfect classification. As we can see in this example the modified trajectories corresponding to $X^{[3]}(t)$ can be classified without error using only one feature. Also the trajectories become more clearly separable after more corrections are applied.

It is possible to observe a similar effect when the noise process is Brownian. In this case, the functions in the corresponding RKHS are the absolutely continuous functions whose evaluation in 0 yields 0 and whose first derivative is in $L^2[0,1]$. A possible example is to use a discontinuous function as the mean, such as a step function with the step at 0.5. In Figure 5.9 an example of that case is shown. Once

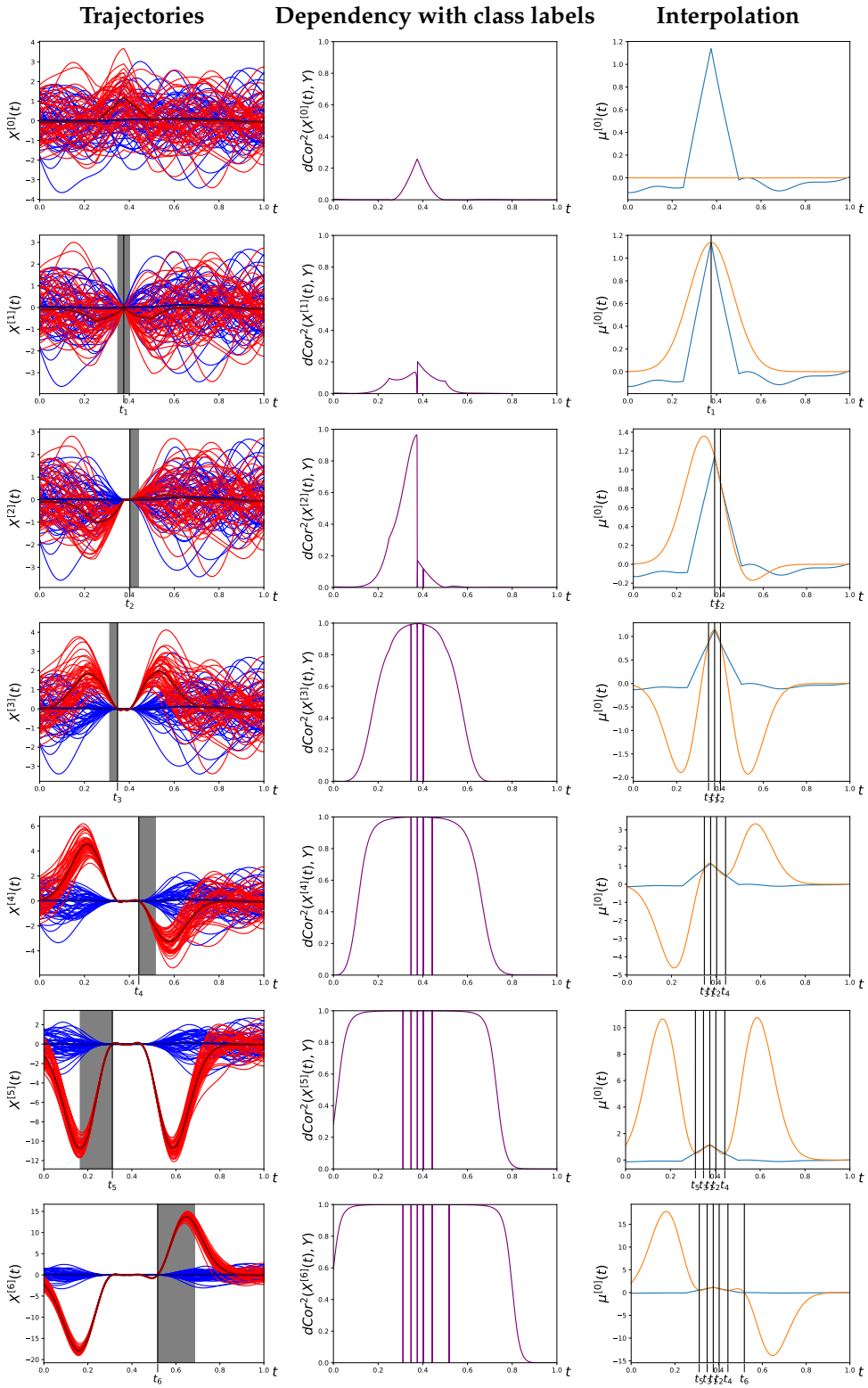

FIGURE 5.8: Example of RMH with a GP correction and a RBF kernel. The noise process has also a RBF kernel and the mean is not in the associated RKHS, so the processes are mutually singular.

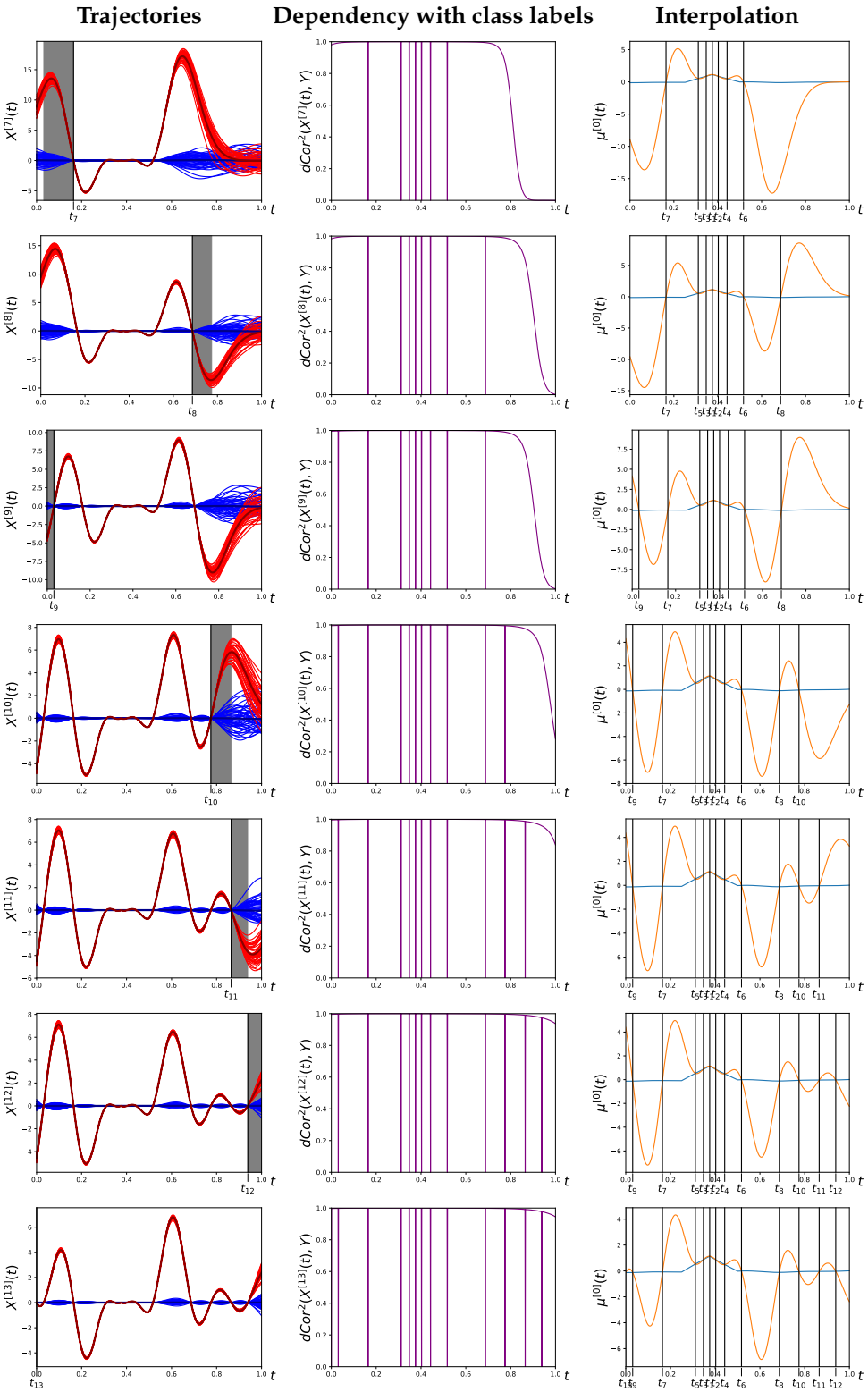

FIGURE 5.8: Example of RMH with a GP correction and a RBF kernel. The noise process has also a RBF kernel and the mean is not in the associated RKHS, so the processes are mutually singular.

again, we can see that only one feature is necessary to achieve a perfect classification for the corrected trajectories corresponding to $X^{[1]}(t)$. However, in this case the next correction makes the two classes indistinguishable.

Another example also using the Brownian kernel is to add a sinusoidal function to the previously used step function, and use this as the mean. Figure 5.10 shows an example of this kind of function. We can see in Figure 5.11 how RMH behave with this mean when the noise process is Brownian. As in the other examples, we can separate both classes using only one feature of the trajectories corresponding to $X^{[2]}(t)$. Also, as in the RBF example, this separation does not disappear after more corrections have been applied.

The causes for this behavior will be investigated in future work. Specifically we want to test in which cases the near-perfect classification phenomenon can be detected applying RMH and looking for features that provide a perfect classification by themselves.

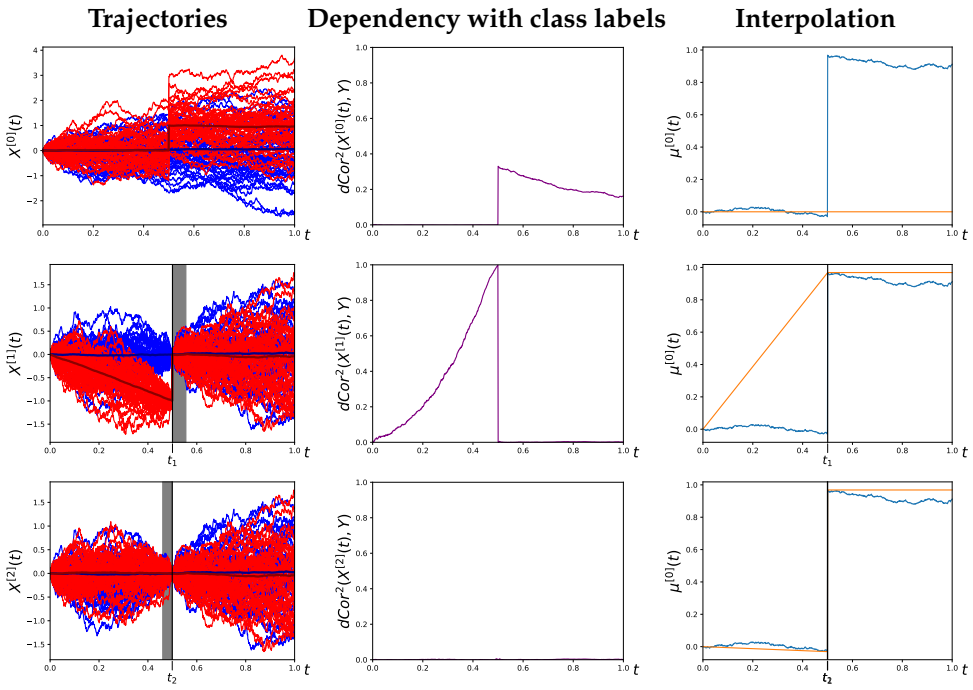

FIGURE 5.9: Example of RMH with a GP correction and a Brownian kernel. The noise process has also a Brownian kernel and the mean is not in the associated RKHS, so the processes are mutually singular.

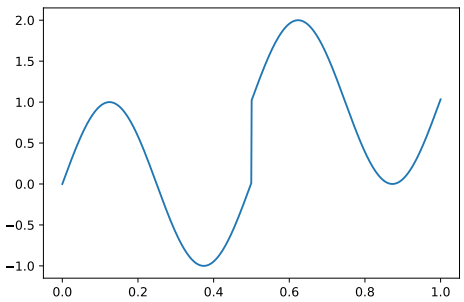

FIGURE 5.10: A step function with the step at 0.5 plus a sinusoidal function. This function does not belong to the RKHS of the Brownian kernel.

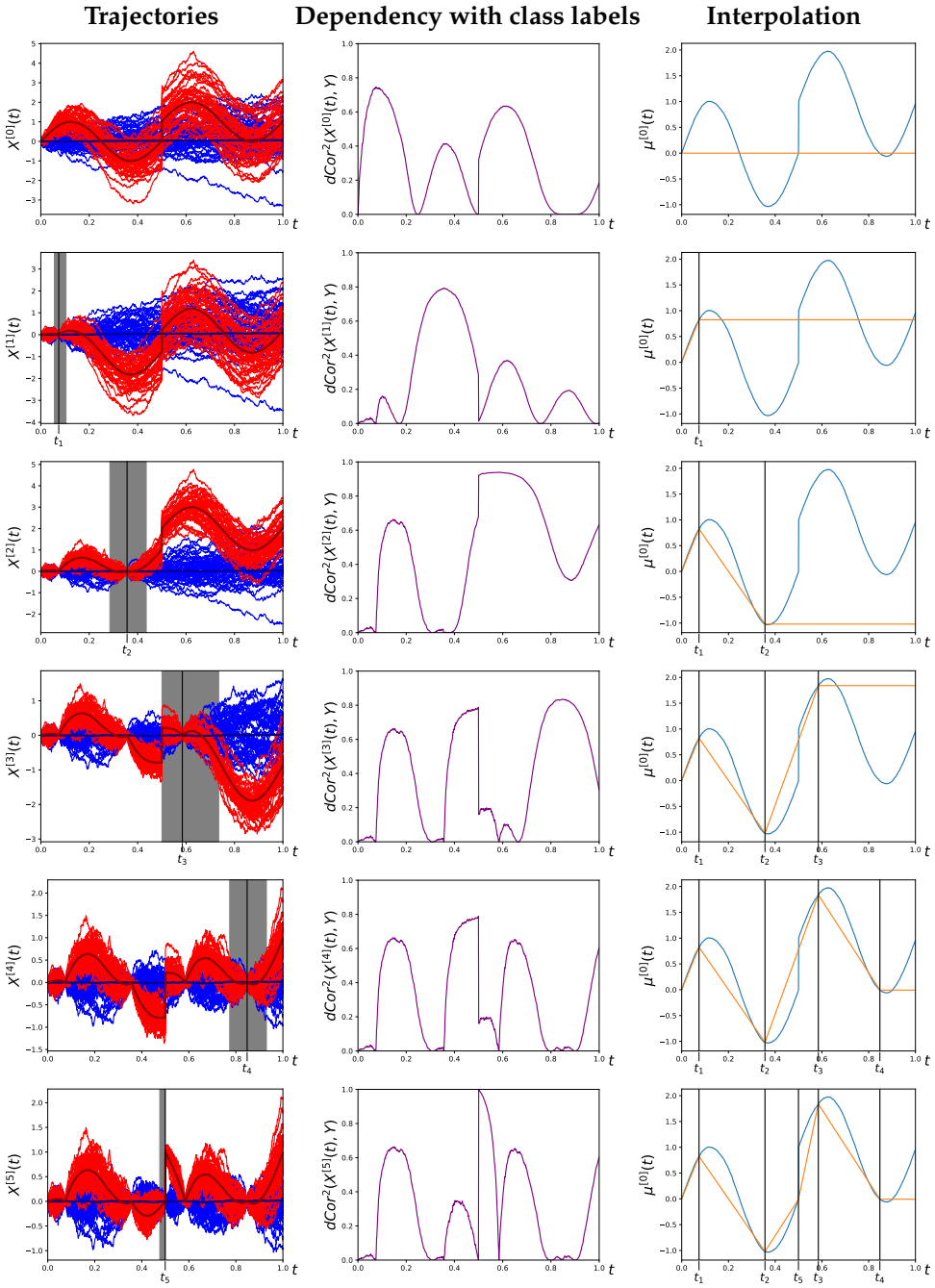

FIGURE 5.11: Example of RMH with a GP correction and a Brownian kernel. The noise process has also a Brownian kernel and the mean is not in the associated RKHS, so the processes are mutually singular.

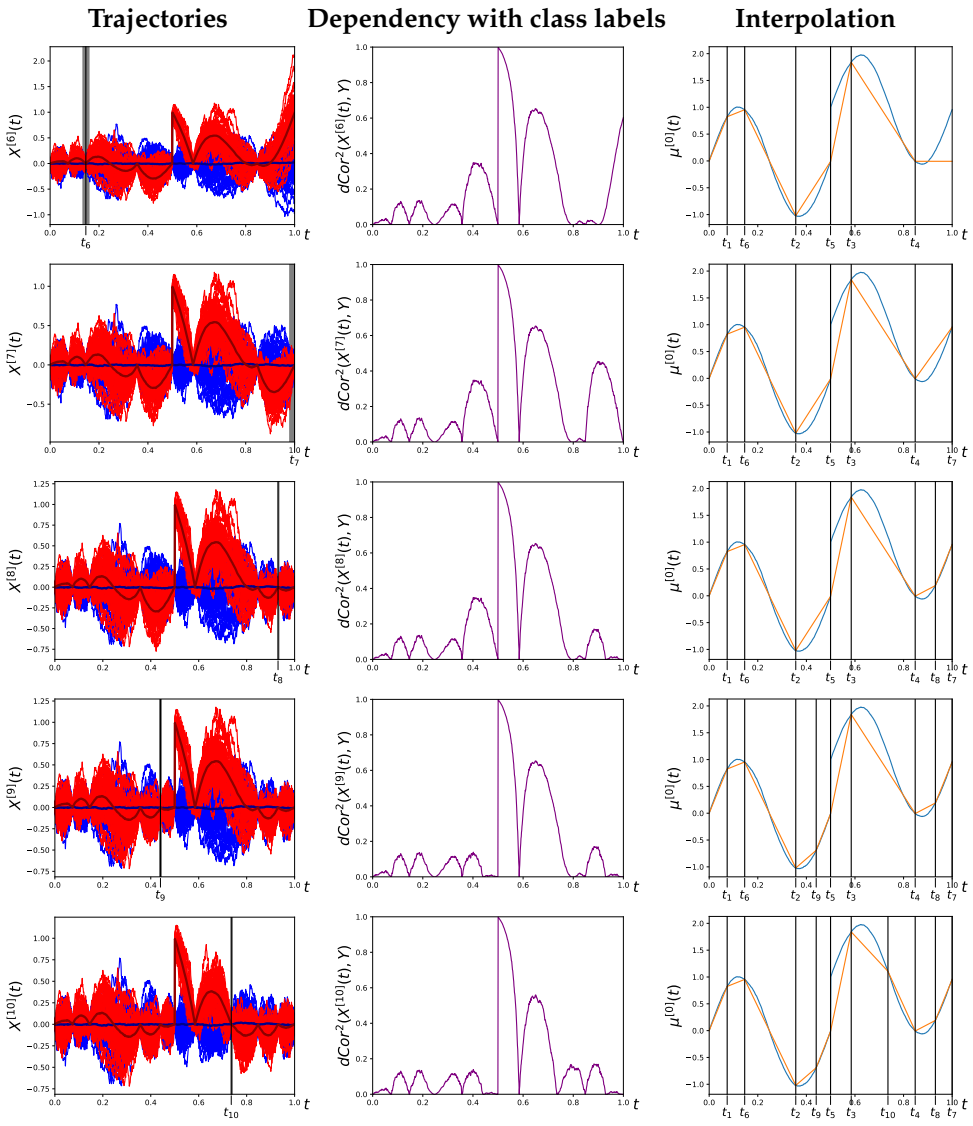

FIGURE 5.11: Example of RMH with a GP correction and a Brownian kernel. The noise process has also a Brownian kernel and the mean is not in the associated RKHS, so the processes are mutually singular.

Chapter 6

Conclusions and future work

We have presented Recursive Maxima Hunting (RMH) with Uniform-Brownian correction, a feature selection method in the context of Functional Data Analysis. We have motivated the use of such methods as a previous step for binary classification, noticing that there exists families of problems where the best possible classification rule depends only on a finite number of points.

Recursive Maxima Hunting selects iteratively the feature that maximizes a measure of dependency with random variable corresponding with the class labels. Then, it removes the information provided by the selected variable from the sampled trajectories. Therefore, the next iterations will select different features. We have shown that, this way, Recursive Maxima Hunting can select variables that are relevant when taken together, even if some of them are not relevant by themselves.

We have proved that RMH can be implemented in an efficient way that only requires to compute the covariances of the original unmodified process.

We have also offered a new perspective on the behaviour of RMH. RMH can be seen as a process that tries to interpolate the difference of the class means between the selected points. This interpolation depends on the nature of the noise stochastic process assumed by RMH to compute the corrections. For example, assuming a Brownian noise process the interpolation is a piecewise linear function beginning at 0. RMH halts when the interpolation becomes an accurate description of this mean difference.

We have proved that, using the Uniform-Brownian correction, RMH can find the points that appear in the optimal classification rule when the mean of one of the classes is zero, the mean of the other is a piecewise linear function, and the noise process is the limit of an Ornstein-Uhlenbeck process when the lengthscale and variance parameters l and σ^2 are in proportion $\sigma^2 = \frac{1}{2}l$ and tend to infinity. We have also show empirically that RMH selects the same features in examples with a different noise process. Therefore, we must assume that the noise process assumed by the RMH method to do the computations has a greater influence in the feature selection process than the real noise. If the mean is not piecewise linear, RMH with the Uniform-Brownian correction will try to build a sufficiently accurate piecewise linear approximation.

Finally, we have obtained empirical data comparing the use of RMH as a first step in classification with other dimensionality reduction methods (and with no dimensionality reduction) in many real and synthetic datasets. We have found that RMH offers a great accuracy and selects few variables in most datasets. However, for some datasets RMH selects variables that, in spite of being relevant, are unnecessary to provide a good classification. This situation arises because RMH is trying to provide an accurate interpolation of the mean difference, even if a more coarse approximation would suffice to classify correctly.

There are a number of issues to resolve and improvement that can be made to the RMH method. Specifically, the stopping conditions for the algorithm need to be explored. In its current formulation, RMH halts when the number of points selected allow one to build a sufficiently accurate interpolation of the difference between the means of the two classes, as shown in section 3.4. However, it is possible that a good classifier can be built using only a subset of those points, without necessarily selecting all of them. One possibility is to measure the dependency between the set of the selected variables and the class to determine whether the algorithm must stop, instead of only the last variable selected, and stop when it does not increase above a threshold.

It is also desirable to better understand the behavior of RMH in problems that present near-perfect classification, which have been analyzed in subsection 5.2.2. We have show that in some examples, exists a feature that unequivocally can determine the class after some corrections have been made. Also, for some examples the number of features with this property increases after more corrections are applied. If we could determine the conditions for this behaviour, it could provide a way to detect these types of problems.

Another development is the extension of RMH to address functional learning problems in which the instances are characterized by vector fields (e.g. panel data), multiclass classification, and regression problems. We will also explore the generalization of RMH to functions that depend on a vector parameter. The presented algorithm can be extended to two or more dimensions using multidimensional Gaussian processes and discarding connected sets of points instead of intervals. We want to implement that extension and test the performance of RMH with images instead of curves.

Another interesting property to explore is the ability of RMH to select a set of suitable points to interpolate a continuous function. This could be used to determine the nodes for a spline or a piecewise linear function that is an accurate approximation to a given continuous function.

Finally, we would like to explore the properties and applications Uniform-Brownian process in further detail. It has some properties (uniformity, stationarity, Markovianity) that make it especially attractive not only in the context of RMH, but also in the general field of FDA. In particular, it could be useful as a reference in computations that involve Radon-Nikodym derivatives. To perform such computations, one would obtain the Radon-Nikodym derivative with respect to the OU process, use this derivative as needed and, eventually, take the large lengthscale, large variance limit, keeping the ratio of these quantities constant.

Appendix A

Properties of Gaussian random vectors and Gaussian processes

This appendix summarizes, for completeness, known results about Gaussian random vectors and Gaussian processes. As a Gaussian process is determined by its marginals, which are Gaussian random vectors, the results for Gaussian random vectors can be also used in the context of Gaussian processes.

Theorem A.1 (Marginal and conditional distributions of a Gaussian)**.** If $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ are random vectors and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ is a normal random vector with mean vector $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and covariance matrix given by

$$\operatorname{Cov}(\boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

then the marginals $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are also normal, with

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}_i] &= \boldsymbol{\mu}_i \\ \operatorname{Cov}(\boldsymbol{Z}_i) &= \boldsymbol{\Sigma}_{ii} \end{aligned}$$

and the conditional distribution of $\boldsymbol{Z}_1 \mid \boldsymbol{Z}_2 = \boldsymbol{x}_2$ is also normal, with

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}_1 \mid \boldsymbol{Z}_2 = \boldsymbol{z}_2] &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{z}_2 - \boldsymbol{\mu}_2) \\ \operatorname{Cov}(\boldsymbol{Z}_1 \mid \boldsymbol{Z}_2 = \boldsymbol{z}_2) &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T. \end{aligned}$$

A proof of this theorem is given on Bishop, 2006.

Theorem A.2 (Linear combination of normal random vector marginals)**.** Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be a Gaussian random vector. Then for any real vector $\boldsymbol{a} = (a_1, \ldots, a_n)$ and any real number b, the linear combination

$$b + \sum_{i=0}^{n} a_i \boldsymbol{Z}_i$$

is a Gaussian random vector.

The proof of the above theorem in the zero mean case, when the $\boldsymbol{Z}_i$ are one dimensional and without independent term can be found on Gallager, 2013. The extension to an arbitrary mean with nonzero independent term is trivial. The case when the $\boldsymbol{Z}_i$ have arbitrary dimensions can be proved using the one dimensional case and the definition of a normal random vector.

Theorem A.3. Let Z be a Gaussian Markov process with covariance function K. Then for all s, t, u with $s < t < u$ it must verify:

$$K(s,u) = \frac{K(s,t)K(t,u)}{K(t,t)}$$

The above theorem was proved in Lamperti, 1977.

Appendix B

Proofs of the theorems

This appendix collects the proofs of the theorems stated in this work.

Proof of Corollary 2.2.2.
Consider the random variables

$$X_0(t) = X(t) \mid Y = 0 \sim N(0, \sigma(t)^2)$$
$$X_1(t) = X(t) \mid Y = 1 \sim N(\mu(t), \sigma(t)^2)$$

for a particular $t \in [0, 1]$.

Let X_0' and $X_1'(t)$ independent copies of $X_0(t)$ and $X_1(t)$ respectively. Then

$$X_0(t) - X_0'(t) \sim N(0 - 0, \sigma(t)^2 + \sigma(t)^2) = N(0, 2\sigma(t)^2)$$
$$X_1(t) - X_1'(t) \sim N(\mu(t) - \mu(t), \sigma(t)^2 + \sigma(t)^2) = N(0, 2\sigma(t)^2)$$
$$X_1(t) - X_0(t) \sim N(\mu(t) - 0, \sigma(t)^2 + \sigma(t)^2) = N(\mu(t), 2\sigma(t)^2)$$

If X is a Gaussian random variable $X \sim N(\mu, \sigma^2)$, then:

$$\mathbb{E}|X| = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left(2\text{cdf}\left(\frac{\mu}{\sigma}\right) - 1\right)$$

Therefore:

$$I_{00}(t) = \mathbb{E}|X_0(t) - X_0'(t)| = \sqrt{2}\sigma(t)\sqrt{\frac{2}{\pi}} = \frac{2\sigma(t)}{\sqrt{\pi}}$$
$$I_{11}(t) = \mathbb{E}|X_1(t) - X_1'(t)| = \sqrt{2}\sigma(t)\sqrt{\frac{2}{\pi}} = \frac{2\sigma(t)}{\sqrt{\pi}}$$
$$\begin{aligned} I_{01}(t) = I_{10}(t) &= \mathbb{E}|X_1(t) - X_0(t)| \\ &= \sqrt{2}\sigma(t)\sqrt{\frac{2}{\pi}}e^{-\frac{\mu(t)^2}{4\sigma(t)^2}} + \mu(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right) - 1\right) \\ &= \frac{2\sigma(t)}{\sqrt{\pi}}e^{-\frac{\mu(t)^2}{4\sigma(t)^2}} + \mu(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right) - 1\right) \end{aligned}$$

We can now apply the following formula (Theorem 2.2.1) to get an expression for the distance covariance:

$$\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2\left[I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2}\right]$$

And we obtain the following expression:

$$\mathcal{V}^2(X(t),Y) = 4p^2(1-p)^2\left[\frac{2\sigma(t)}{\sqrt{\pi}}\left(e^{-\frac{\mu(t)^2}{4\sigma(t)^2}}-1\right)+\mu(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma(t)}\right)-1\right)\right] \tag{B.1}$$

The first and second derivatives of $\mathcal{V}$ are easy to compute, using the relation

$$\frac{d}{dt}\text{cdf}(x) = \text{pdf}(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$

□

Proof of Theorem 3.3.1, Theorem 3.3.2 and Theorem 3.3.3.
Theorem 3.3.1 holds, by the definition of the problem, when $i = 0$.

Since a Gaussian process is defined by its marginal distributions at a finite number of locations, which are Gaussian vectors, if $Z^{[i-1]}$ is a Gaussian zero-mean stochastic process we can compute the conditional distribution of a Gaussian vector for each of its marginals, given the selected point, arriving at the formula in Theorem 3.3.2. This can be seen using Theorem A.1 where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, $\boldsymbol{z}_2 = X^{[i-1]}(t_i)$ and

$$\begin{aligned}\boldsymbol{\Sigma}_{12} &= K^{[i-1]}(t,t_i)\\ \boldsymbol{\Sigma}_{22} &= K^{[i-1]}(t_i,t_i).\end{aligned}$$

If Theorem 3.3.1 is valid for some value of i, then $Z^{[i]}$ is a zero-mean Gaussian process. Therefore the above reasoning applies. This is true when $i = 0$. We can prove that if Theorem 3.3.1 holds for some value of $i-1$, then it holds for i, and by induction Theorem 3.3.1 and Theorem 3.3.2 hold for every value of i.

Suppose that Theorem 3.3.1 holds for $i-1$. Then, Theorem 3.3.2 also holds for $i-1$ and we can write

$$\begin{aligned}X^{[i]}(t) &= X^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = X^{[i-1]}(t_i)\right]\\ &= X^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}X^{[i-1]}(t_i).\end{aligned}$$

If we consider that the trajectories have class 1 we can replace $X^{[i-1]}$ by its expression in Theorem 3.3.1 and one obtains

$$\begin{aligned}X^{[i]}(t) &= \mu^{[i-1]}(t) + Z^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}\left(\mu^{[i-1]}(t_i) + Z^{[i-1]}(t_i)\right)\\ &= \left(\mu^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}\mu^{[i-1]}(t_i)\right) + \left(Z^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}Z^{[i-1]}(t_i)\right)\\ &= \mu^{[i]}(t) + Z^{[i]}(t)\end{aligned}$$

where

$$\begin{aligned}\mu^{[i]}(t) &= \mu^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}\mu^{[i-1]}(t_i)\\ &= \mu^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i) = \mu^{[i-1]}(t_i)\right]\\ Z^{[i]}(t) &= Z^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}Z^{[i-1]}(t_i)\\ &= Z^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right].\end{aligned}$$

For class 0 trajectories, the same derivation can be made using $\mu^{[i-1]}(t) = 0$. Therefore the mean after the correction is still 0 and the two classes have the same noise process. This proves Theorem 3.3.1.

Let us now will prove that $Z^{[i]}(t)$ is Gaussian. Starting from the form of the corrected noise process

$$Z^{[i]}(t) = Z^{[i-1]}(t) - \frac{K^{[i-1]}(t,t_i)}{K^{[i-1]}(t_i,t_i)}Z^{[i-1]}(t_i).$$

Since $Z^{[i-1]}(t)$ is Gaussian, for every set of points $\{s_1,\dots,s_n\}$ the joint distribution of $Z^{[i-1]}(s_1),\dots,Z^{[i-1]}(s_n)$ and $Z^{[i-1]}(t_i)$ is a multivariate Gaussian. Therefore, $(Z^{[i]}(s_1),\dots,Z^{[i]}(s_n))$ is a linear combination of two Gaussian processes whose joint distribution is Gaussian. Thus, by Theorem A.2, it is a Gaussian vector. Since Gaussian processes are defined by their marginals, and every marginal of $Z^{[i]}(t)$ is Gaussian, then $Z^{[i]}(t)$ is a Gaussian process.

Finally, we prove that $Z^{[i]}(t)$ has mean 0.

$$\begin{aligned}\mathbb{E}\left[Z^{[i]}(t)\right] &= \mathbb{E}\left[Z^{[i-1]}(t) - \mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right]\right]\\ &= \mathbb{E}\left[Z^{[i-1]}(t)\right] - \mathbb{E}\left[\mathbb{E}\left[Z^{[i-1]}(t) \mid Z^{[i-1]}(t_i)\right]\right]\\ &= \mathbb{E}\left[Z^{[i-1]}(t)\right] - \mathbb{E}\left[Z^{[i-1]}(t)\right]\\ &= 0\end{aligned}$$

The second equality is a consequence of the linearity of the expectation. The third one is the law of total expectation.

We have then proved that if Theorem 3.3.1 holds for $i-1$ then it also holds for i. As said before we can apply induction to show that the three theorems hold for every value of i. □

Proof of Lemma 3.3.6.
Using Theorem 3.3.1, the left hand side process is Gaussian. As a conditioned Gaussian process is also Gaussian, the right hand side is a Gaussian process too. Thus, it suffices to show that the expectations and covariance functions for both processes are the same.

Let

$$\boldsymbol{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \quad \Sigma_{\boldsymbol{rs}} = \mathrm{Cov}(Z(\boldsymbol{r}), Z(\boldsymbol{s})), \quad \text{and } Z(\boldsymbol{r}) = \begin{pmatrix} Z(r_1) \\ \vdots \\ Z(r_i) \end{pmatrix},$$

for arbitrary vectors

$$\boldsymbol{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_i \end{pmatrix} \text{ and } \boldsymbol{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_j \end{pmatrix}.$$

The expectation of $Z(t) - \mathbb{E}\left[Z(t) \mid Z(t_1) \dots Z(t_n)\right]$ is

$$\begin{aligned} \mathbb{E}\left[Z - \mathbb{E}\left[Z \mid Z(t_1) \dots Z(t_n)\right]\right] &= \mathbb{E}\left[Z\right] - \mathbb{E}\left[\mathbb{E}\left[Z \mid Z(t_1) \dots Z(t_n)\right]\right] \\ &= \mathbb{E}\left[Z\right] - \mathbb{E}\left[Z\right] = 0 \end{aligned}$$

where we have used the linearity of the expectation and the law of total expectation.

The expectation of $\left[Z(t) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right]$ can be computed using the formula for the conditional distribution of Gaussian vectors (Theorem A.1) with

$$\begin{aligned} \boldsymbol{\mu}_1 &= \mathbb{E}[Z(s)], \\ \boldsymbol{\mu}_2 &= \mathbb{E}[Z(\boldsymbol{t})], \\ \boldsymbol{z}_2 &= Z(\boldsymbol{t}), \\ \boldsymbol{\Sigma}_{12} &= \Sigma_{s\boldsymbol{t}}, \\ \boldsymbol{\Sigma}_{22} &= \Sigma_{\boldsymbol{tt}}, \end{aligned}$$

and recalling that the formula is also valid for Gaussian processes. Then, the expectation is

$$\begin{aligned} \mathbb{E}\left[Z(s) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right] &= \mathbb{E}[Z(s)] + \Sigma_{s\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}(Z(\boldsymbol{t}) - \mathbb{E}[Z(\boldsymbol{t})]) \\ &= 0 + \Sigma_{s\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}(0 - 0) = 0. \end{aligned}$$

Since $\mathbb{E}[Z(s)] = 0$ and $\mathbb{E}[Z(\boldsymbol{t})] = 0$,

$$Z(s) - \mathbb{E}\left[Z(s) \mid Z(t_1) \dots Z(t_n)\right] = \begin{pmatrix} 1 & -\Sigma_{s\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1} \end{pmatrix} \begin{pmatrix} Z(s) \\ Z(\boldsymbol{t}) \end{pmatrix},$$

where $\Sigma_{s\boldsymbol{t}}$ is a $1 \times n$ matrix and $\Sigma_{\boldsymbol{tt}}^{-1}$ is a $n \times n$ matrix. Then, its covariance will be

$$\begin{aligned} &\operatorname{Cov}(Z(r) - \mathbb{E}\left[Z(r) \mid Z(\boldsymbol{t})\right], Z(s) - \mathbb{E}\left[Z(s) \mid Z(\boldsymbol{t})\right]) \\ &= \mathbb{E}\left[\begin{pmatrix} I & -\Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1} \end{pmatrix} \begin{pmatrix} Z(r) \\ Z(\boldsymbol{t}) \end{pmatrix} \begin{pmatrix} Z(s) & [Z(\boldsymbol{t})]^T \end{pmatrix} \begin{pmatrix} I \\ -\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} \end{pmatrix}\right] \\ &= \begin{pmatrix} I & -\Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1} \end{pmatrix} \mathbb{E}\left[\begin{pmatrix} Z(r) \\ Z(\boldsymbol{t}) \end{pmatrix} \begin{pmatrix} Z(s) & [Z(\boldsymbol{t})]^T \end{pmatrix}\right] \begin{pmatrix} I \\ -\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} \end{pmatrix} \\ &= \begin{pmatrix} I & -\Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_{rs} & \Sigma_{r\boldsymbol{t}} \\ \Sigma_{\boldsymbol{t}s} & \Sigma_{\boldsymbol{tt}} \end{pmatrix} \begin{pmatrix} I \\ -\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} \end{pmatrix} \\ &= \Sigma_{rs} - \Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} - \Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} + \Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{tt}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s} \\ &= \Sigma_{rs} - \Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s}. \end{aligned}$$

The covariance function of $\left[Z(t) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right]$ can be computed using Theorem A.1, as well:

$$\begin{aligned} &\operatorname{Cov}\left(\left[Z(r) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right], \left[Z(s) \mid Z(t_1) = 0 \dots Z(t_n) = 0\right]\right) \\ &= \Sigma_{rs} - \Sigma_{r\boldsymbol{t}}\Sigma_{\boldsymbol{tt}}^{-1}\Sigma_{\boldsymbol{t}s}. \end{aligned}$$

Since the means and covariance functions are the same, and the two processes are Gaussian, then they are the same process. □

Proof of Theorem 3.3.4.
We want to prove that

$$Z^{[i]}(t) = \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0 \dots Z^{[0]}(t_i) = 0 \right]$$

by induction on i.

We know that the formula is true for $i = 0$:

$$Z^{[0]}(t) = Z^{[0]}(t)$$

We suppose that the formula is true for $i = n - 1$:

$$Z^{[n-1]}(t) = \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0 \dots Z^{[0]}(t_{n-1}) = 0 \right]$$

Then, we have

$$\begin{aligned} Z^{[n]}(t) &= Z^{[n-1]}(t) - \mathbb{E}\left[Z^{[n-1]}(t) \mid Z^{[n-1]}(t_n) \right] \\ &= \left[Z^{[n-1]}(t) \mid Z^{[n-1]}(t_n) = 0 \right] \\ &= \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0, \dots, Z^{[0]}(t_n) = 0 \right], \end{aligned}$$

where we have used Lemma 3.3.6 in the intermediate step. In the last step we have used the induction hypothesis, together with the fact that the conditioning of a stochastic process on the value of several random variables can be done also iteratively. That is, for every stochastic process Z and every points r, s, if we denote $Y(t) = Z(t) \mid Z(r)$ then

$$[Y(t) \mid Y(s)] = [Z(t) \mid Z(r), Z(s)] .$$

□

Proof of Theorem 3.3.5.
We want to prove

$$\mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1), \dots Z^{[0]}(t_n) \right] = \sum_{i=0}^{n-1} \mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1}) \right]$$

As $\mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1}) \right]$ is the formula for the $(i+1)$-th correction, we have that

$$Z^{[0]} - \sum_{i=0}^{n-1} \mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1}) \right] = Z^{[n]}$$

because $Z^{[n]}$ is the process after n corrections have been made.

Theorem 3.3.4 implies that

$$\begin{aligned} Z^{[n]}(t) &= \left[Z^{[0]}(t) \mid Z^{[0]}(t_1) = 0 \dots Z^{[0]}(t_n) = 0\right] \\ &= Z^{[0]}(t) - \mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1) \dots Z^{[0]}(t_n)\right] \end{aligned}$$

where, in the last step, we have used Lemma 3.3.6.

Combining the two expressions we have

$$Z^{[0]} - \sum_{i=0}^{n-1} \mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1})\right] = Z^{[0]}(t) - \mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1) \dots Z^{[0]}(t_n)\right],$$

and thus

$$\sum_{i=0}^{n-1} \mathbb{E}\left[Z^{[i]}(t) \mid Z^{[i]}(t_{i+1})\right] = \mathbb{E}\left[Z^{[0]}(t) \mid Z^{[0]}(t_1) \dots Z^{[0]}(t_n)\right],$$

as we wanted to prove. □

Proof of Theorem 3.3.7.
From Theorem 3.3.4 and Theorem A.1, we have

$$K_1(s,u) = K(s,u) - \frac{K(s,t)K(t,u)}{K(t,t)}$$

From Theorem A.3 we have:

$$\begin{aligned} K_1(s,u) &= K(s,u) - \frac{K(s,t)K(t,u)}{K(t,t)} \\ &= K(s,u) - K(s,u) \\ &= 0 \end{aligned}$$

□

Proof of Theorem 3.4.1.
Lets define:

$$\boldsymbol{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \Sigma_{\boldsymbol{rs}} = \mathrm{Cov}(Z(\boldsymbol{r}), Z(\boldsymbol{s})), \quad \text{and } Z(\boldsymbol{r}) = \begin{pmatrix} Z(r_1) \\ \vdots \\ Z(r_i) \end{pmatrix},$$

for arbitrary vectors

$$\boldsymbol{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_i \end{pmatrix} \text{ and } \boldsymbol{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_j \end{pmatrix}.$$

By Theorem A.1 we know that:

$$\begin{aligned} \mathbb{E}\left[Z(t) \mid Z(t_1) = \mu_1, \dots Z(t_n) = \mu_n\right] &= \mathbb{E}[Z(t)] + \boldsymbol{\Sigma}_{\boldsymbol{tt}} \boldsymbol{\Sigma}_{\boldsymbol{tt}}^{-1} (\boldsymbol{\mu} - \mathbb{E}[Z(\boldsymbol{t})]) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{tt}} \boldsymbol{\Sigma}_{\boldsymbol{tt}}^{-1} \boldsymbol{\mu}. \end{aligned}$$

(i) is easy to prove. If $t = t_i$ then $\boldsymbol{\Sigma}_{\boldsymbol{tt}}$ is the i-th row of $\boldsymbol{\Sigma}_{\boldsymbol{tt}}$, which we denote $[\boldsymbol{\Sigma}_{\boldsymbol{tt}}]_{i*}$.

As

$$\boldsymbol{\Sigma_{tt}}\boldsymbol{\Sigma_{tt}}^{-1} = I_n$$

we have that:

$$\boldsymbol{\Sigma_{tt}}\boldsymbol{\Sigma_{tt}}^{-1} = [\boldsymbol{\Sigma_{tt}}]_{i*}\,\boldsymbol{\Sigma_{tt}}^{-1} = [I_n]_{i*}\,.$$

Then:

$$\begin{aligned}
&\mathbb{E}\left[Z(t) \mid Z(t_1) = \mu_1, \dots Z(t_n) = \mu_n\right] \\
&= \boldsymbol{\Sigma_{tt}}\boldsymbol{\Sigma_{tt}}^{-1}\boldsymbol{\mu} \\
&= [I_n]_{i*}\,\boldsymbol{\mu} \\
&= \mu_i.
\end{aligned}$$

For the following parts, since Z is a Brownian process, $\mathrm{Cov}(Z(r), Z(s)) = K(r,s) = \min(r,s)$. Therefore

$$\boldsymbol{\Sigma_{tt}} = \begin{pmatrix} \min(t,t_1) & \dots & \min(t,t_n) \end{pmatrix}$$

$$\boldsymbol{\Sigma_{tt}} = \begin{pmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & \dots & t_n \end{pmatrix}.$$

If $t = t_i$, then:

$$\boldsymbol{\Sigma}_{t_i\boldsymbol{t}} = \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & \dots & t_i \end{pmatrix}.$$

Lets prove (ii). Suppose that $t \in (t_i, t_{i+1})$. Then:

$$\begin{aligned}
\boldsymbol{\Sigma_{tt}} &= \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t & \dots & t \end{pmatrix} \\
&= \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t_i \dots & t_i \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 & 0 & t-t_i \dots & t-t_i \end{pmatrix} \\
&= \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t_i \dots & t_i \end{pmatrix} + \frac{t-t_i}{t_{i+1}-t_i} \begin{pmatrix} 0 & \dots & 0 & 0 & t_{i+1}-t_i \dots & t_{i+1}-t_i \end{pmatrix} \\
&= \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t_i \dots & t_i \end{pmatrix} \\
&\quad + \frac{t-t_i}{t_{i+1}-t_i} \left(\begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t_{i+1} \dots & t_{i+1} \end{pmatrix} - \begin{pmatrix} t_1 & \dots & t_{i-1} & t_i & t_i \dots & t_i \end{pmatrix} \right) \\
&= \boldsymbol{\Sigma}_{t_i\boldsymbol{t}} + \frac{t-t_i}{t_{i+1}-t_i} \left(\boldsymbol{\Sigma}_{t_{i+1}\boldsymbol{t}} - \boldsymbol{\Sigma}_{t_i\boldsymbol{t}} \right).
\end{aligned}$$

Now applying the distributive property of matrix product we have:

$$\begin{aligned}
&\mathbb{E}\left[Z(t) \mid Z(t_1) = \mu_1, \dots Z(t_n) = \mu_n\right] \\
&= \boldsymbol{\Sigma_{tt}}\boldsymbol{\Sigma_{tt}}^{-1}\boldsymbol{\mu} \\
&= \boldsymbol{\Sigma}_{t_i\boldsymbol{t}}\boldsymbol{\Sigma_{tt}}^{-1}\boldsymbol{\mu} + \frac{t-t_i}{t_{i+1}-t_i} \left(\boldsymbol{\Sigma}_{t_{i+1}\boldsymbol{t}}\boldsymbol{\Sigma_{tt}}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}_{t_i\boldsymbol{t}}\boldsymbol{\Sigma_{tt}}^{-1}\boldsymbol{\mu} \right) \\
&= \mu_i + \frac{t-t_i}{t_{i+1}-t_i}(\mu_{i+1} - \mu_i) \\
&= \mu_i \left(1 - \frac{t-t_i}{t_{i+1}-t_i} \right) + \mu_{i+1} \left(\frac{t-t_i}{t_{i+1}-t_i} \right),
\end{aligned}$$

which is the result we set out to prove.

In order to prove (iii), it is easy to see that $f(0) = 0$, because in this case $\boldsymbol{\Sigma}_{0\boldsymbol{t}} = \mathbf{0}$. If $t \in (0, t_1)$ we can repeat the previous demonstration:

$$\begin{aligned}
\boldsymbol{\Sigma_{tt}} &= \begin{pmatrix} t & \dots & t \end{pmatrix} \\
&= \frac{t}{t_1} \begin{pmatrix} t_1 & \dots & t_1 \end{pmatrix} \\
&= \frac{t}{t_1} \boldsymbol{\Sigma}_{t_1\boldsymbol{t}}.
\end{aligned}$$

Thus:

$$\begin{aligned}
&\mathbb{E}\left[Z(t) \mid Z(t_1) = \mu_1, \dots Z(t_n) = \mu_n\right] \\
&= \boldsymbol{\Sigma}_{t\boldsymbol{t}} \boldsymbol{\Sigma}_{\boldsymbol{tt}}^{-1} \boldsymbol{\mu} \\
&= \frac{t}{t_1} \boldsymbol{\Sigma}_{t_1 \boldsymbol{t}} \boldsymbol{\Sigma}_{\boldsymbol{tt}}^{-1} \boldsymbol{\mu} \\
&= \mu_1 \left(\frac{t}{t_1}\right).
\end{aligned}$$

(iv) is trivial to prove, because if $t > t_n$ then:

$$\boldsymbol{\Sigma}_{t\boldsymbol{t}} = \boldsymbol{\Sigma}_{t_n \boldsymbol{t}}.$$

□

Proof of Theorem 4.1.1.
X verifies the conditions of Corollary 2.2.2. Therefore we have that, for σ constant

$$\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[\frac{2\sigma}{\sqrt{\pi}}\left(e^{-\frac{\mu(t)^2}{4\sigma^2}} - 1\right) + \mu(t)\left(2\text{cdf}\left(\frac{\mu(t)}{\sqrt{2}\sigma}\right) - 1\right)\right],$$

where $p = \mathbb{P}[Y = 1]$.

$\mathcal{V}^2(X(t), Y)$ depends on t only through μ. Therefore it can be written as a function of μ:

$$f(\mu) = 4p^2(1-p)^2 \left[\frac{2\sigma}{\sqrt{\pi}}\left(e^{-\frac{\mu^2}{4\sigma^2}} - 1\right) + \mu\left(2\text{cdf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) - 1\right)\right]$$

Now, we can differentiate this function to obtain:

$$\begin{aligned}
f'(\mu) &= 4p^2(1-p)^2 \left[\frac{2\sigma}{\sqrt{\pi}} e^{-\frac{\mu^2}{4\sigma^2}}\left(-\frac{\mu}{2\sigma^2}\right) + \left(2\text{cdf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) - 1\right) + 2\mu\text{pdf}\left(\frac{\mu}{\sigma\sqrt{2}}\right)\frac{1}{\sigma\sqrt{2}}\right] \\
&= 4p^2(1-p)^2 \left[-\frac{\mu}{\sigma\sqrt{\pi}} e^{-\frac{\mu^2}{4\sigma^2}} + 2\text{cdf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) - 1 + \frac{\mu}{\sigma\sqrt{\pi}} e^{-\frac{\mu^2}{4\sigma^2}}\right] \\
&= 4p^2(1-p)^2 \left[2\text{cdf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) - 1\right]
\end{aligned}$$

Since $\text{cdf}(x) > 0.5$ if $x > 0$ and $\text{cdf}(x) = 1 - \text{cdf}(-x)$ we have that $f'(\mu) > 0$ if $\mu > 0$ and also $f'(\mu) = -f'(-\mu)$. Thus $f(\mu)$ is a monotonically increasing function of $|\mu|$. It follows that $\mathcal{V}^2(X(t), Y) = f(\mu(t))$ has the same increasing and decreasing intervals as $|\mu(t)|$.

□

Proof of Theorem 4.1.2.
X verifies the conditions of Corollary 2.2.2. Therefore, for μ constant

$$\mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \left[\frac{2\sigma(t)}{\sqrt{\pi}}\left(e^{-\frac{\mu^2}{4\sigma(t)^2}} - 1\right) + \mu\left(2\text{cdf}\left(\frac{\mu}{\sqrt{2}\sigma(t)}\right) - 1\right)\right],$$

where $p = \mathbb{P}[Y = 1]$. If $\mu = 0$ then the two terms are zero and so $\mathcal{V}^2(X(t), Y) = 0$.

Otherwise, $\mathcal{V}^2(X(t), Y)$ depends on t, only through σ.

$$g(\sigma) = 4p^2(1-p)^2 \left[\frac{2\sigma}{\sqrt{\pi}}\left(e^{-\frac{\mu^2}{4\sigma^2}} - 1\right) + \mu\left(2\text{cdf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) - 1\right)\right]$$

The derivative of this function with respect to σ is

$$\begin{aligned} g'(\sigma) = 4p^2(1-p)^2 \Bigg[& \frac{2}{\sqrt{\pi}} \left(e^{-\frac{\mu^2}{4\sigma^2}} - 1 \right) + \frac{2\sigma}{\sqrt{\pi}} e^{-\frac{\mu^2}{4\sigma^2}} \left(-\frac{\mu^2}{4} \right) (-2) \left(\frac{1}{\sigma^3} \right) \\ & + 2\mu \text{pdf} \left(\frac{\mu}{\sigma\sqrt{2}} \right) \frac{\mu}{\sqrt{2}} (-1) \frac{1}{\sigma^2} \Bigg] \\ = 4p^2(1-p)^2 \Bigg[& \frac{2}{\sqrt{\pi}} \left(e^{-\frac{\mu^2}{4\sigma^2}} - 1 \right) + \frac{\mu^2}{\sqrt{\pi}\sigma^2} e^{-\frac{\mu^2}{4\sigma^2}} - \frac{\mu^2}{\sqrt{\pi}\sigma^2} e^{-\frac{\mu^2}{4\sigma^2}} \Bigg] \\ = 4p^2(1-p)^2 \Bigg[& \frac{2}{\sqrt{\pi}} \left(e^{-\frac{\mu^2}{4\sigma^2}} - 1 \right) \Bigg]. \end{aligned}$$

Since $g'(\sigma) < 0, \quad \forall \sigma > 0$, $g(\sigma)$ is a monotonically decreasing function of σ. It follows that $\mathcal{V}^2(X(t), Y) = g(\sigma(t))$ increases where $\sigma(t)$ decreases and vice versa. □

Proof of Theorem 4.2.1.
Assume that $\mathcal{V}^2(X(t), Y)$ has a maximum at t_0. If $\mu(t_0) = 0$, by Corollary 2.2.2 $\mathcal{V}^2(X(t_0), Y) = 0$, which is the minimum possible value of the distance covariance. Since t_0 is a maximum, then $\mathcal{V}^2(X(t), Y) = 0$ for every t.

Consider now the case $\mu(t_0) \neq 0$. Assuming that $\mathcal{V}^2(X(t), Y)$ is twice differentiable at $t = t_0$, then its first derivative at this point is 0 and its second derivative is negative. By Corollary 2.2.2, the second derivative is

$$\begin{aligned} \frac{d^2}{dt^2} \mathcal{V}^2(X(t), Y) = 4p^2(1-p)^2 \Bigg[& \frac{2\sigma''(t)}{\sqrt{\pi}} \left(e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} - 1 \right) \\ & + \frac{1}{\sqrt{\pi}} e^{-\frac{\mu(t)^2}{4\sigma^2(t)}} \left(\frac{(\mu(t)\sigma'(t) - \sigma(t)\mu'(t))^2}{\sigma^3(t)} \right) \\ & + \mu''(t) \left(2\text{cdf} \left(\frac{\mu(t)}{\sqrt{2}\sigma(t)} \right) - 1 \right) \Bigg] \end{aligned}$$

Since $\sigma''(t) < 0$ and $\mu(t) \neq 0$, for every t, the first term is always positive. The second term is positive or zero. The third term is zero because $\mu''(t) = 0$. Thus the second derivative is positive and the point t_0 can not be a maximum. □

Proof of Theorem 4.3.1.
Since Z is Gaussian Markov, from Theorem A.3, for all s, t, u with $s < t < u$:

$$K(s,u) = \frac{K(s,t)K(t,u)}{K(t,t)}.$$

Since Z is stationary, $K(s,u)$ can be rewritten as a function of $|s-u|$ so:

$$K(|s-u|) = \frac{K(|s-t|)K(|t-u|)}{K(0)}.$$

Given that $\sigma^2 = K(0)$,

$$K(|s-t|+|t-u|) = \frac{K(|s-t|)K(|t-u|)}{\sigma^2}.$$

If $f(x+y) = f(x)f(y)$ and f is continuous, then $f(x) = e^{kx}$ with k constant. If we define $l = -\frac{1}{k}$, then K has to be of the form

$$K(s,t) = \sigma^2 \exp\left(-\frac{|s-u|}{l}\right)$$

as we wanted to prove. The condition $l > 0$ ensures that K is positive semidefinite. □

Appendix C

Plots of the experiments with real datasets

In this appendix, the plots of the datasets used in subsection 5.1.2 are shown, along with box plots illustrating the error and number of variables selected for each method in the dataset.

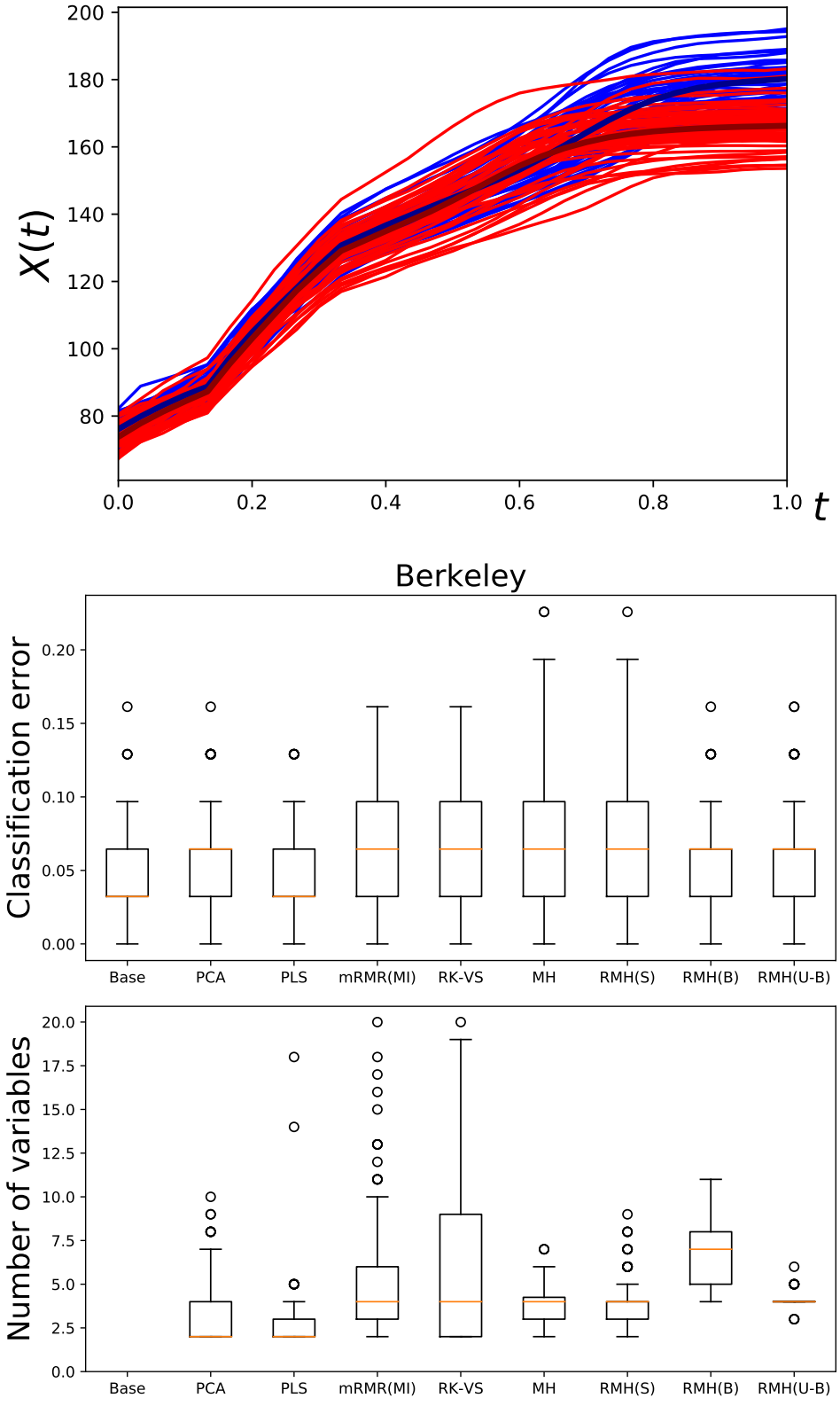

FIGURE C.1: The first figure show the trajectories in the Berkeley Growth dataset. The two box plots correspond to the classification error and the number of variables selected.

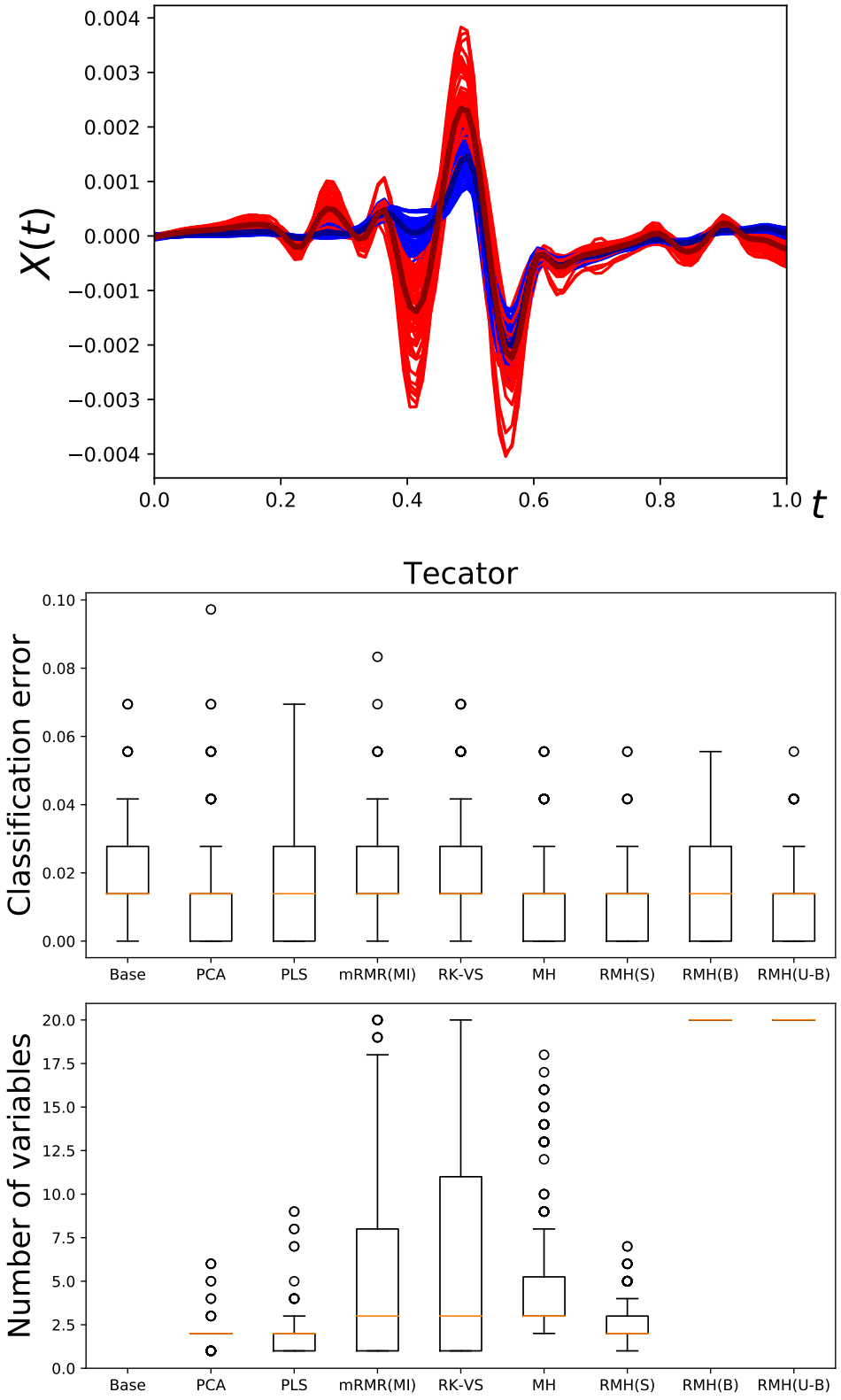

FIGURE C.2: The first figure show the trajectories in the Tecator dataset. The two box plots correspond to the classification error and the number of variables selected.

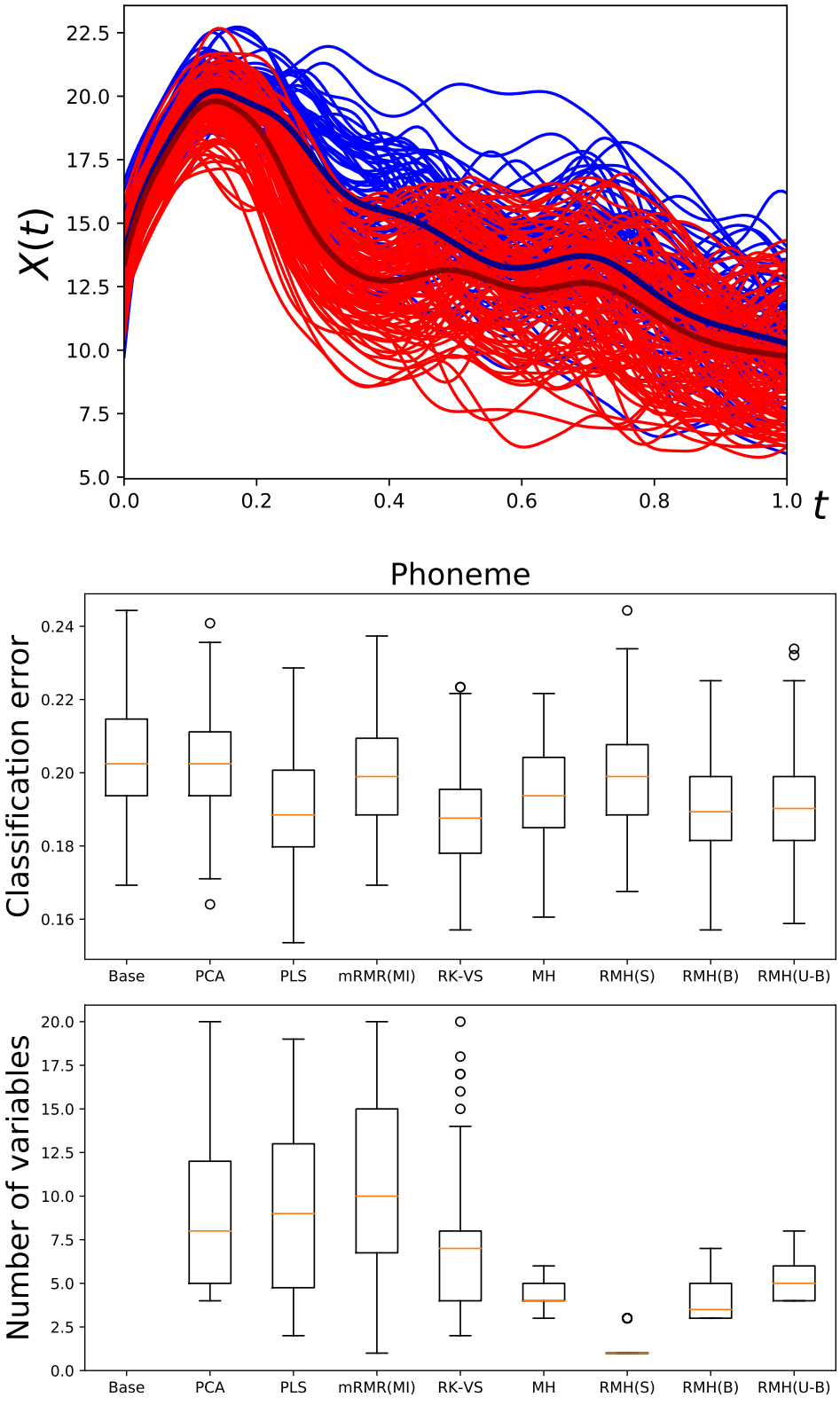

FIGURE C.3: The first figure show the trajectories in the Phoneme dataset. The two box plots correspond to the classification error and the number of variables selected.

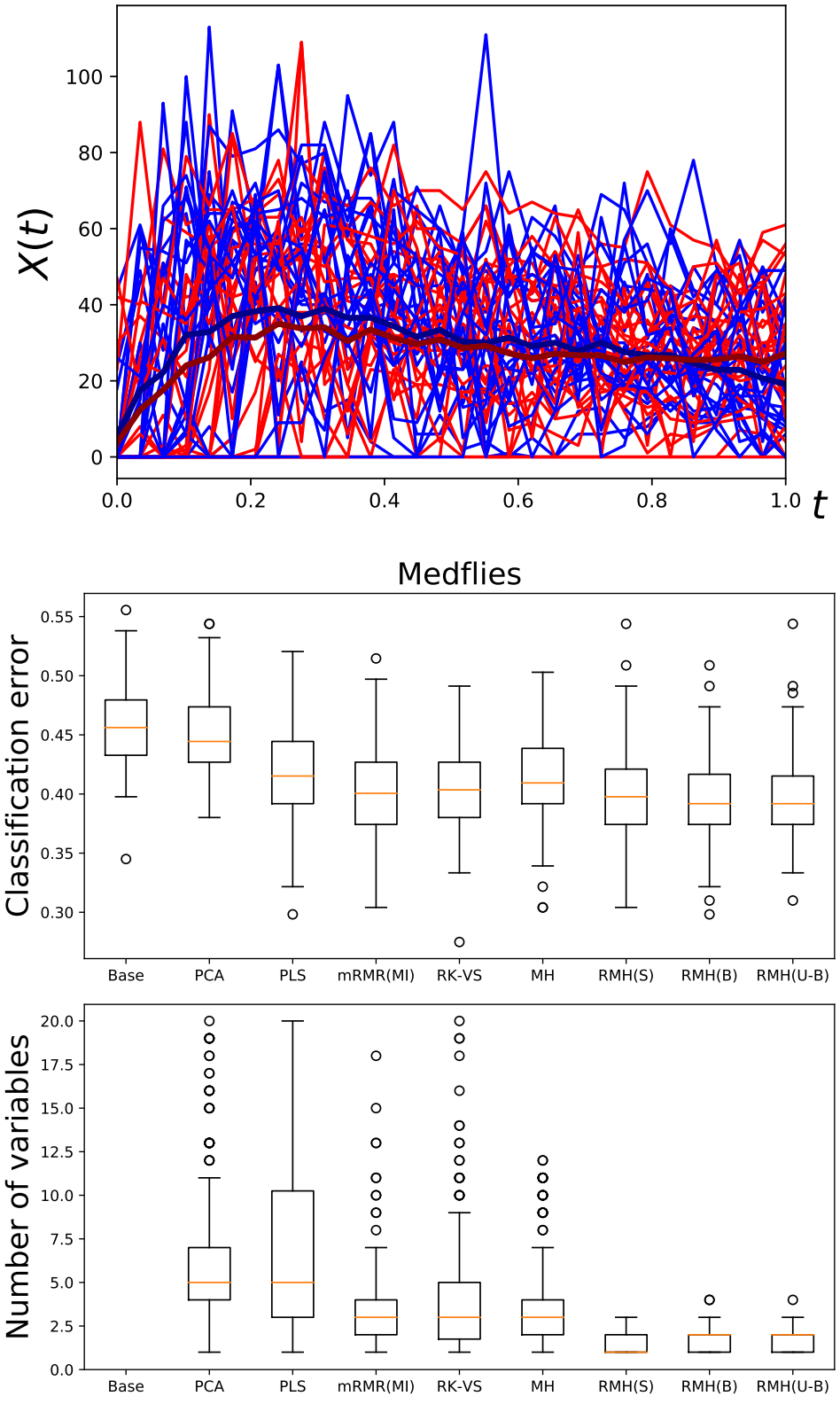

FIGURE C.4: The first figure show the trajectories in the Medflies dataset. The two box plots correspond to the classification error and the number of variables selected.

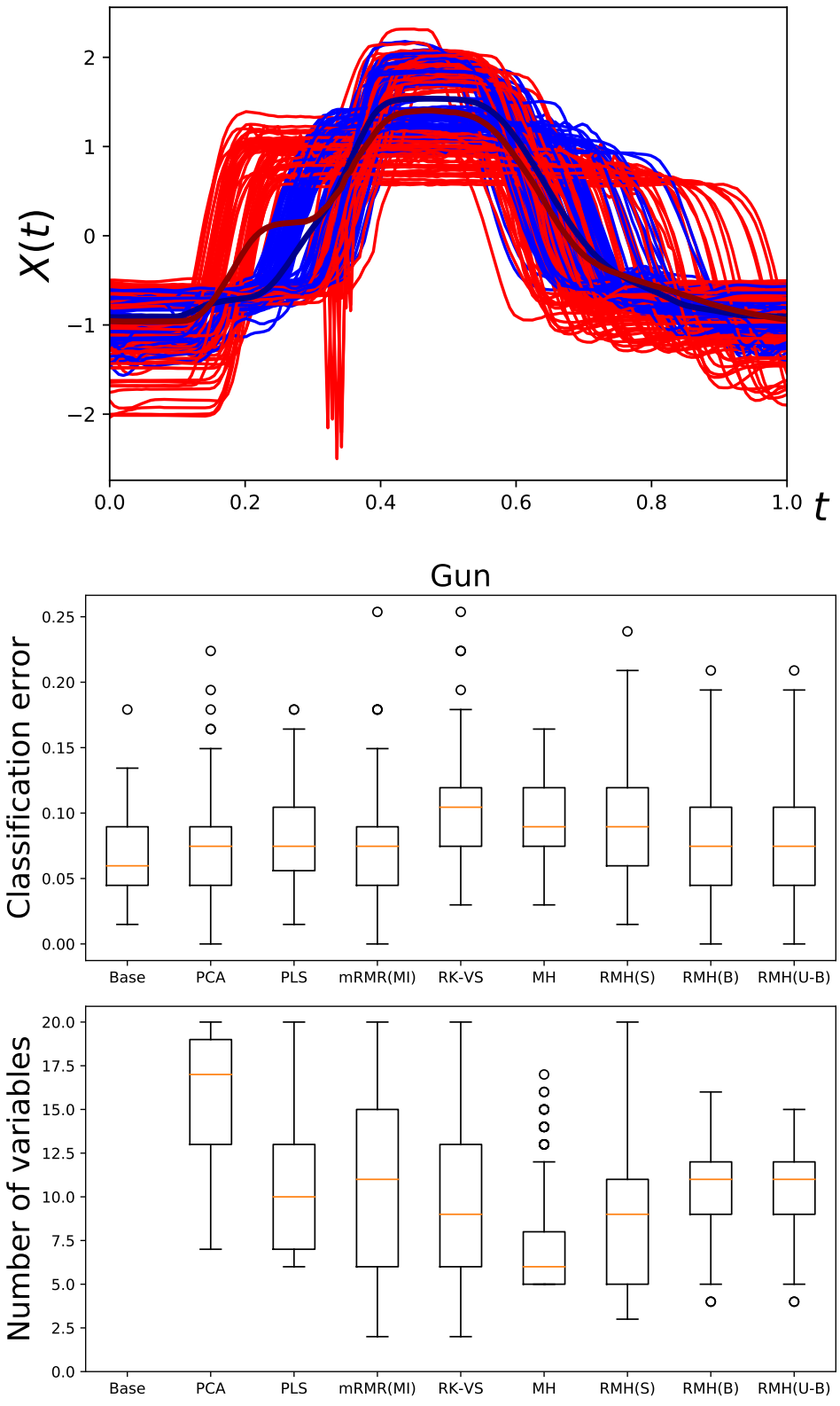

FIGURE C.5: The first figure show the trajectories in the Gun dataset. The two box plots correspond to the classification error and the number of variables selected.

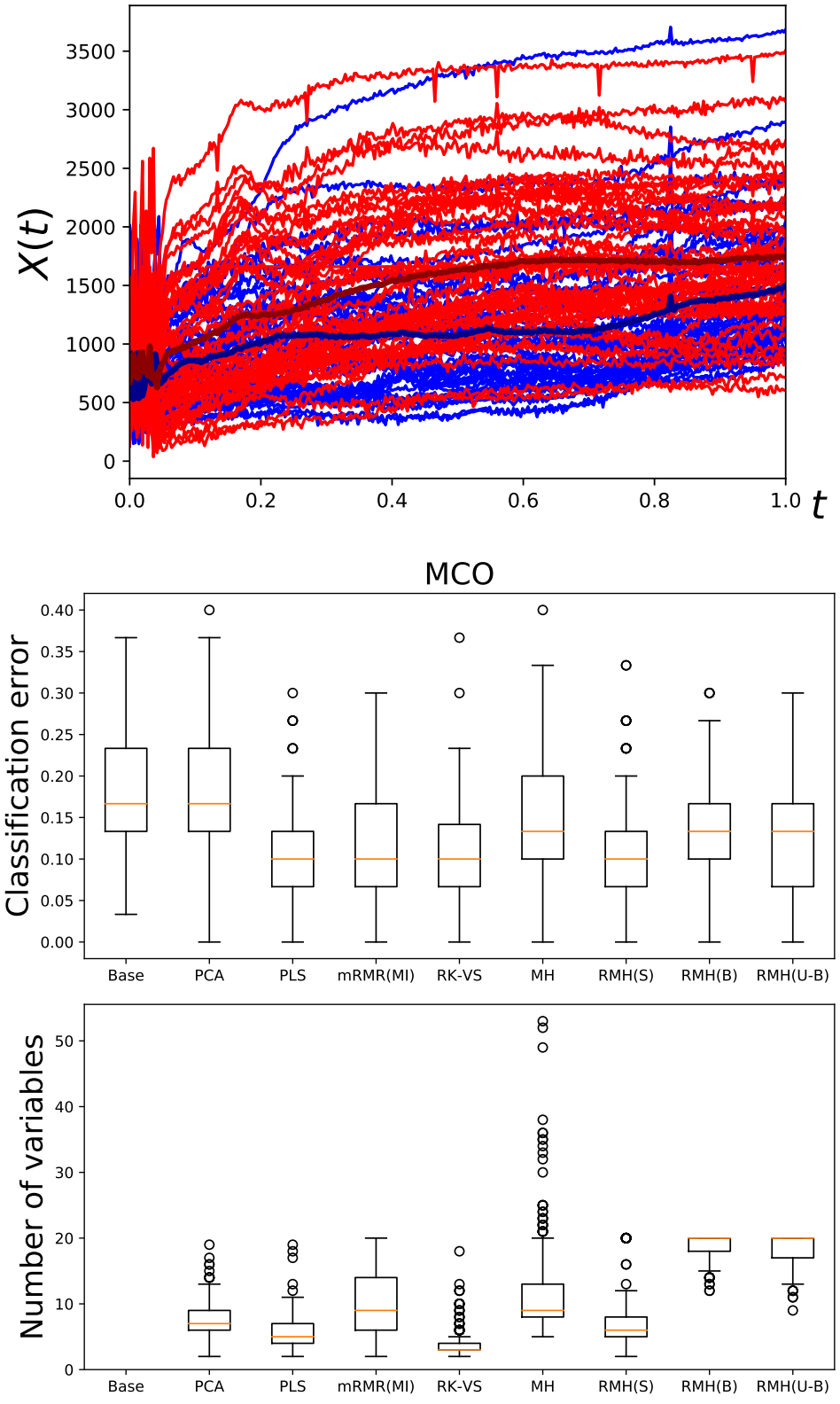

FIGURE C.6: The first figure show the trajectories in the MCO dataset. The two box plots correspond to the classification error and the number of variables selected.

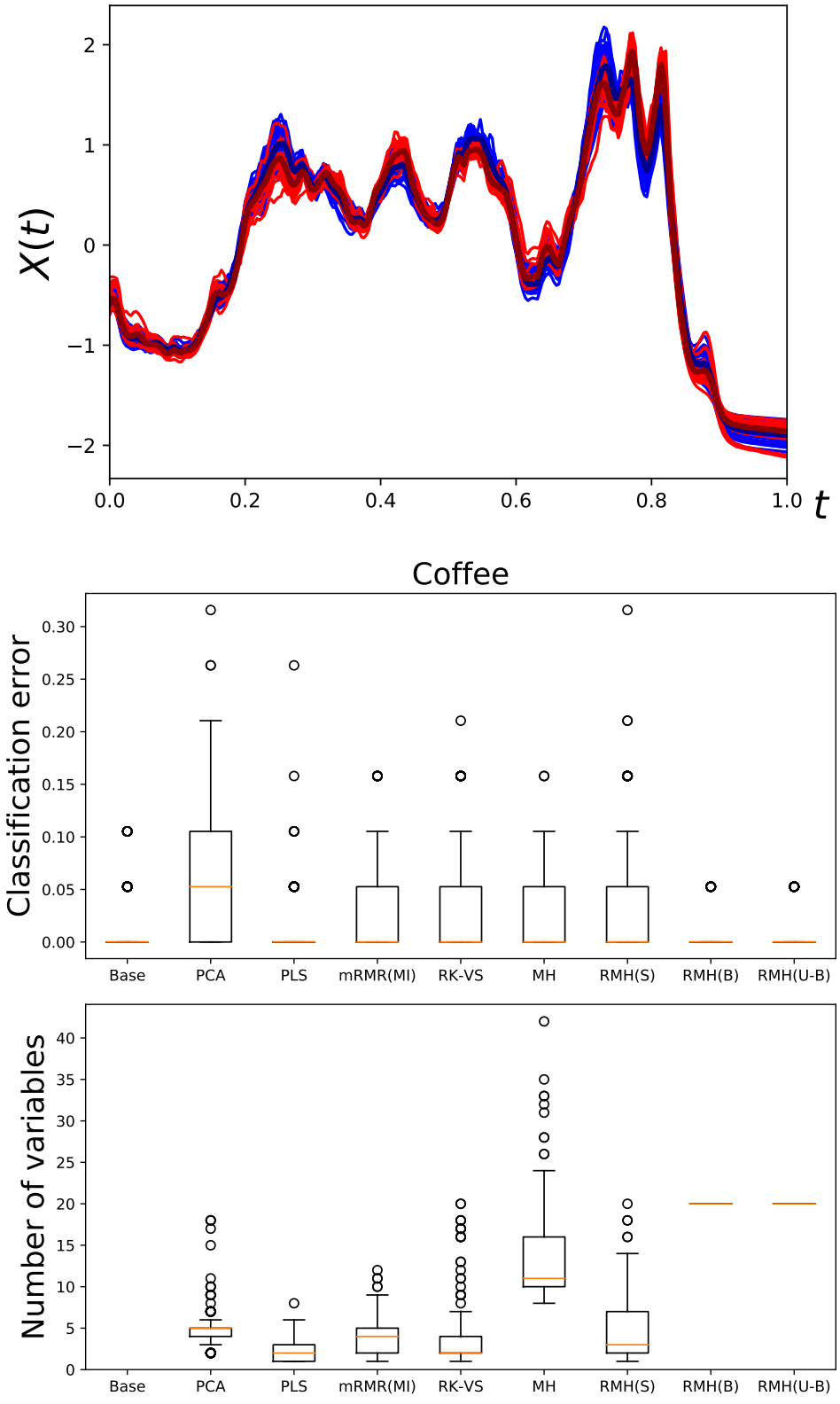

FIGURE C.7: The first figure show the trajectories in the Coffee dataset. The two box plots correspond to the classification error and the number of variables selected.

Appendix D

Comparison between RMH with the real and the sample covariance

In the experiments shown in section 5.1, we have included as a method RMH using as the covariance function the covariance matrix obtained from the sample data, and assuming that the noise process is Gaussian and thus the correction formula still applies. One could expect that the sample covariance will have the same performance in synthetic data and better performance in real datasets, but that does not coincide with the observed behavior.

We have tested the method in a small experiment, subtracting, for each trajectory, the mean of its class before computing the sample covariance. We have made an experiment with synthetic data using a Brownian noise process and a piecewise linear mean of the same kind as the ones studied in section 2.1.2. The relevant points are those in with two straight lines intersect. In Figure D.1 we can see that RMH using a Brownian correction selects the right points.

In Figure D.2 it is shown the results using as the covariance function the sample covariance. We can see that the results are similar to the real one as expected, but as the estimated covariance is still noticeably different from the real one, the first correction does not leave every trajectory beginning at the origin. That causes numerical errors, and as a result, the algorithm mistakes the origin for a relevant point, and selects it. This explains why, in the experiments with synthetic data, this correction always selected at least one more point than the Brownian correction.

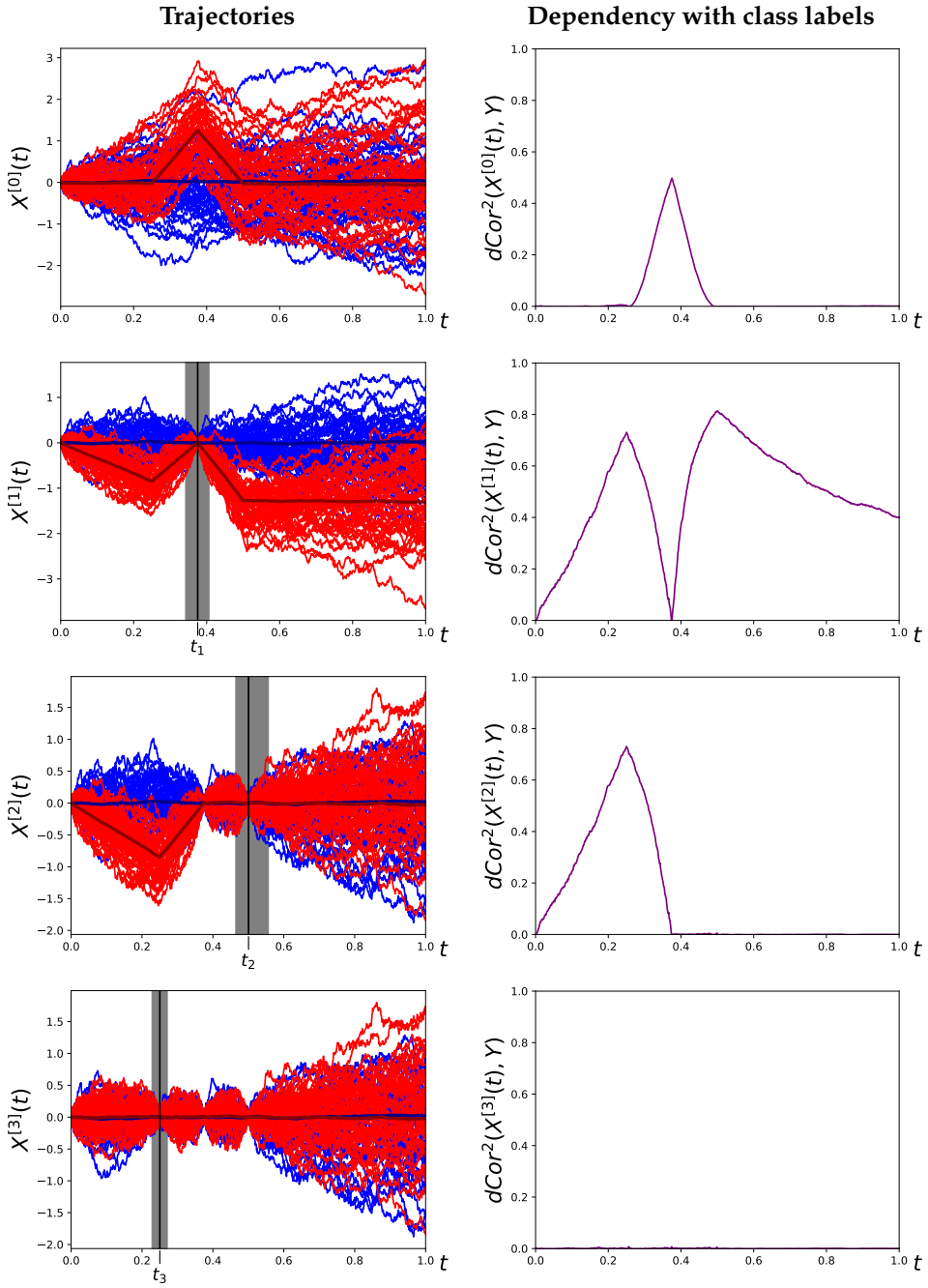

FIGURE D.1: Example of the execution of RMH with Brownian trajectories and where the nonzero mean is a simple piecewise linear function with a peak shape.

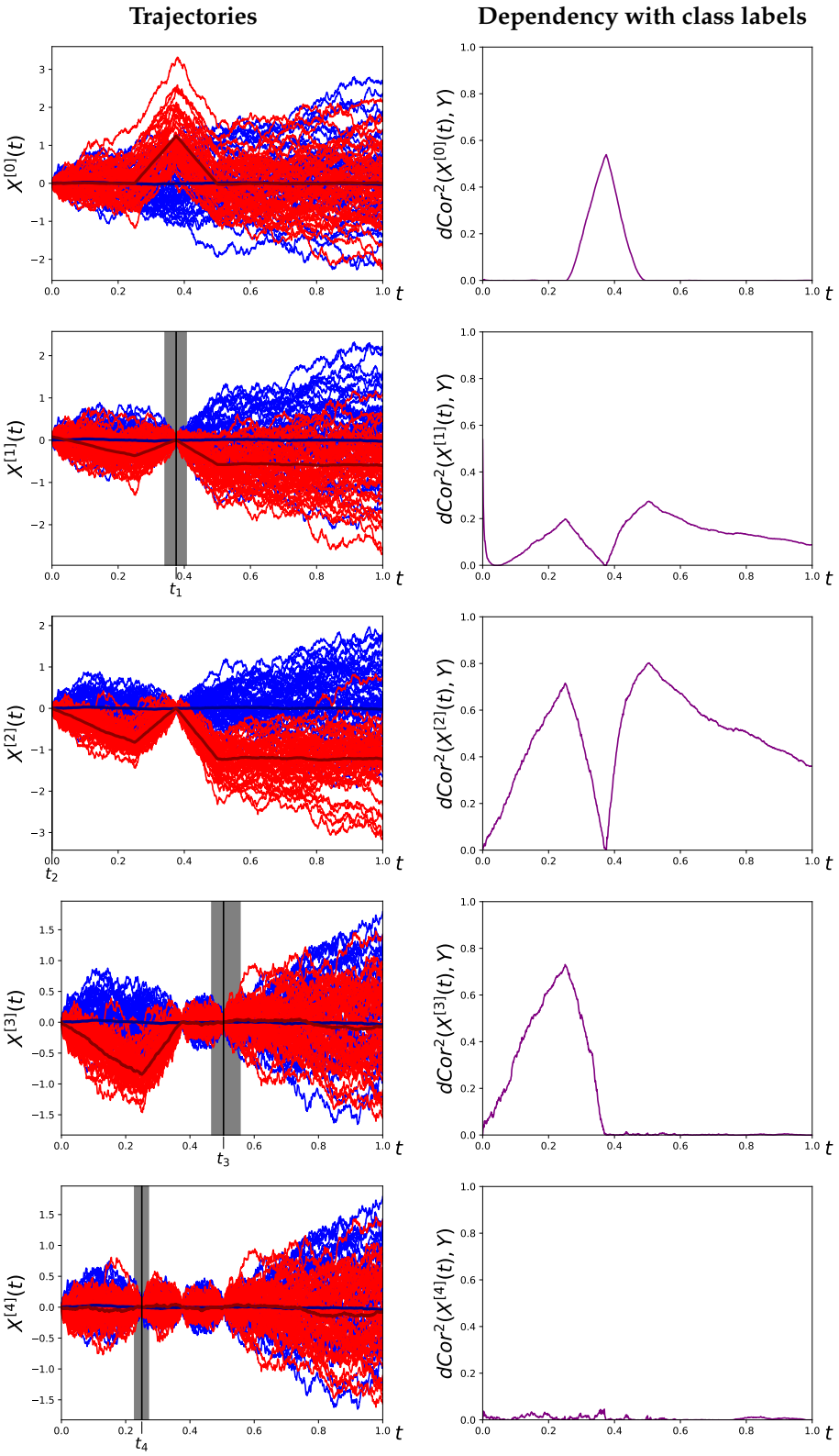

FIGURE D.2: Example of the execution of RMH with Brownian trajectories with the same means as Figure D.1, whose covariance has been estimated using the sample, after the means of each class have been subtracted.

Appendix E

Kernel list

This appendix list the kernels (covariance functions) for several Gaussian processes used throughout this work.

Brownian kernel

The kernel corresponding to a Brownian process is

$$K(s,t) = \sigma^2 \min(s,t),$$

where $\sigma^2 > 0$ is the *variance* at point 1, $K(1,1)$. For a standard Brownian process $\sigma^2 = 1$

$$K(s,t) = \min(s,t).$$

Brownian bridge kernel

The kernel corresponding to a Brownian bridge process is

$$K(s,t) = \sigma^2 \frac{(t_2 - s)(t - t_1)}{t_2 - t_1},$$

where t_1 and t_2 are the fixed points and $\sigma^2 > 0$ is a scale parameter.

Exponential kernel

The kernel corresponding to an Ornstein-Uhlenbeck process is the exponential kernel

$$K(s,t) = \sigma^2 \exp\left(-\frac{|s-t|}{l}\right),$$

where $\sigma^2 > 0$ is the *variance* at point 0, $K(0,0)$, and $l > 0$ is the *lengthscale* parameter.

RBF kernel

The kernel corresponding to a RBF process is

$$K(s,t) = \sigma^2 \exp\left(-\frac{|s-t|^2}{2l^2}\right),$$

where $\sigma^2 > 0$ is the *variance* at point 0, $K(0,0)$, and $l > 0$ is the *lengthscale* parameter.

Matern 3/2 kernel

The kernel corresponding to an Matern $3/2$ process is

$$K(s,t) = \sigma^2 \left(1 + \sqrt{3}\frac{|s-t|}{l}\right) \exp\left(-\frac{\sqrt{3}|s-t|}{l}\right),$$

where $\sigma^2 > 0$ is the *variance* at point 0, $K(0,0)$, and $l > 0$ is the *lengthscale* parameter.

Bibliography

Aronszajn, N. (1950). "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68.3, pp. 337–404. ISSN: 00029947. URL: http://www.jstor.org/stable/1990404.

Bagnall, Anthony et al. (2012). "Transformation based ensembles for time series classification". In: *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, pp. 307–318.

Baíllo, Amparo, Antonio Cuevas, and Juan Antonio Cuesta-Albertos (2011). "Supervised Classification for a Family of Gaussian Functional Models". In: *Scandinavian Journal of Statistics* 38.3, pp. 480–498. ISSN: 1467-9469. DOI: 10.1111/j.1467-9469.2011.00734.x.

Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Berrendero, José R., Antonio Cuevas, and José L. Torrecilla (2016a). "The mRMR variable selection method: a comparative study for functional data". In: *Journal of Statistical Computation and Simulation* 86.5, pp. 891–907. DOI: 10.1080/00949655.2015.1042378.

— (2016b). "Variable selection in functional data classification: a maxima-hunting proposal." In: *Statistica Sinica* 26, pp. 619–638. DOI: 10.5705/ss.202014.0014.

— (2017). "On the use of reproducing kernel Hilbert spaces in functional classification". In: *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2017.1320287.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Briandet, Romain, E. Katherine Kemsley, and Reginald H. Wilson (1996). "Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics". In: *Journal of agricultural and food chemistry* 44.1, pp. 170–174.

Butte, Atul J and Isaac S Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Pacific Symposium on Biocomputing*. Vol. 5. 415, p. 26.

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons.

Cuevas, Antonio (2014). "A partial overview of the theory of statistics with functional data". In: *Journal of Statistical Planning and Inference* 147, pp. 1 –23. ISSN: 0378-3758. DOI: http://dx.doi.org/10.1016/j.jspi.2013.04.002.

Cuevas, Antonio, Manuel Febrero, and Ricardo Fraiman (2004). "An anova test for functional data". In: *Computational Statistics & Data Analysis* 47.1, pp. 111 –122. ISSN: 0167-9473. DOI: http://dx.doi.org/10.1016/j.csda.2003.10.021.

Cuevas, Antonio, Manuel Febrero, and Ricardo Fraiman (2006). "On the use of the bootstrap for estimating functions with functional data". In: *Computational Statistics & Data Analysis* 51.2, pp. 1063 –1074. ISSN: 0167-9473. DOI: http://dx.doi.org/10.1016/j.csda.2005.10.012.

Delaigle, Aurore and Peter Hall (2012). "Achieving near perfect classification for functional data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2, pp. 267–286. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.01003.x.

Ding, Chris and Hanchuan Peng (2005). "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of Bioinformatics and Computational Biology* 03.02, pp. 185–205. DOI: 10.1142/S0219720005001004.

Feldman, Jacob (1958). "Equivalence and perpendicularity of Gaussian processes". In: *Pacific Journal of Mathematics* 8.4, pp. 699–708.

Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Galeano, Pedro, Esdras Joseph, and Rosa E. Lillo (2015). "The Mahalanobis Distance for Functional Data With Applications to Classification". In: *Technometrics* 57.2, pp. 281–291. DOI: 10.1080/00401706.2014.902774.

Gallager, Robert G. (2013). *Stochastic processes: theory for applications*. Cambridge University Press, pp. 113–114.

Gretton, Arthur (2013). "Introduction to rkhs, and some simple kernel algorithms". In: *Advanced Topics in Machine Learning. Lecture Conducted from University College London*.

Gulgezen, Gokhan, Zehra Cataltepe, and Lei Yu (2009). "Stable and Accurate Feature Selection". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I*. Ed. by Wray Buntine et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 455–468. ISBN: 978-3-642-04180-8. DOI: 10.1007/978-3-642-04180-8_47.

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Hotelling, Harold (1933). "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6, p. 417.

Huo, Xiaoming and Gábor J. Székely (2016). "Fast Computing for Distance Covariance". In: *Technometrics* 58.4, pp. 435–447. DOI: 10.1080/00401706.2015.1054435.

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger (2004). "Estimating mutual information". In: *Physical Review E* 69 (6), p. 066138. DOI: 10.1103/PhysRevE.69.066138.

Lamperti, John (1977). *Stochastic processes: a survey of the mathematical theory*. Vol. 23. Springer Verlag, New York Inc.

Michaels, George S et al. (1998). "Cluster analysis and data visualization of large-scale gene expression data". In: *Pacific symposium on biocomputing*. Vol. 3, pp. 42–53.

Moon, Young-Il, Balaji Rajagopalan, and Upmanu Lall (1995). "Estimation of mutual information using kernel density estimators". In: *Physical Review E* 52 (3), pp. 2318–2321. DOI: 10.1103/PhysRevE.52.2318.

Mosler, Karl and Pavlo Mozharovskyi (2015). "Fast DD-classification of functional data". In: *Statistical Papers*, pp. 1–35.

Parzen, Emanuel (1961). "An Approach to Time Series Analysis". In: *The Annals of Mathematical Statistics* 32.4, pp. 951–989. ISSN: 00034851. URL: http://www.jstor.org/stable/2237900.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Preda, Cristian, Gilbert Saporta, and Caroline Lévéder (2007). "PLS classification of functional data". In: *Computational Statistics* 22.2, pp. 223–235. ISSN: 1613-9658. DOI: 10.1007/s00180-007-0041-4.

Ramsay, J. O. (1982). "When the data are functions". In: *Psychometrika* 47.4, pp. 379–396. ISSN: 1860-0980. DOI: 10.1007/BF02293704.

Ramsay, James O (2006). *Functional data analysis*. Wiley Online Library.

Ratanamahatana, Chotirat Ann and Eamonn Keogh (2004). "Everything you know about dynamic time warping is wrong". In: *Third Workshop on Mining Temporal and Sequential Data*. Citeseer, pp. 22–25.

Rosipal, Roman and Nicole Krämer (2006). "Overview and recent advances in partial least squares". In: *Lecture notes in computer science* 3940, p. 34.

Roulston, Mark S (1999). "Estimating the errors on measured entropy and mutual information". In: *Physica D: Nonlinear Phenomena* 125.3, pp. 285 –294. ISSN: 0167-2789. DOI: http://dx.doi.org/10.1016/S0167-2789(98)00269-3.

Ruiz-Meana, Marisol et al. (2003). "Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocytes during ischemic conditions". In: *American Journal of Physiology - Heart and Circulatory Physiology* 285.3, H999–H1006. ISSN: 0363-6135. DOI: 10.1152/ajpheart.00035.2003.

Shannon, C. E. (2001). "A Mathematical Theory of Communication". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 5.1, pp. 3–55. ISSN: 1559-1662. DOI: 10.1145/584091.584093.

Steinwart, I., D. Hush, and C. Scovel (2006). "An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels". In: *IEEE Transactions on Information Theory* 52.10, pp. 4635–4643. ISSN: 0018-9448. DOI: 10.1109/TIT.2006.881713.

Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6, pp. 2769–2794. DOI: 10.1214/009053607000000505.

Torrecilla, José L. (2015). "On the theory and practice of variable selection for functional data". PhD thesis. Universidad Autónoma de Madrid.

Torrecilla, José L. and Alberto Suárez (2016). "Feature selection in functional data classification with recursive maxima hunting". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., pp. 4835–4843. URL: http://papers.nips.cc/paper/6392-feature-selection-in-functional-data-classification-with-recursive-maxima-hunting.pdf.

Uhlenbeck, G. E. and L. S. Ornstein (1930). "On the Theory of the Brownian Motion". In: *Phys. Rev.* 36 (5), pp. 823–841. DOI: 10.1103/PhysRev.36.823.

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller (2016). "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3.1, pp. 257–295. DOI: 10.1146/annurev-statistics-041715-033624.

Wegelin, Jacob A et al. (2000). "A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case". In: *University of Washington, Department of Statistics, Tech. Rep*.